UNIVERSITY OF Southampton

CENTRO ECOLOGIA APLICADA "Prof. Baeta Neves"

UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCES AND MATHEMATICS

School of Civil Engineering and the Environment

Centre for Environmental Sciences

in collaboration with

UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR DE AGRONOMIA

Centro de Ecologia Aplicada "Prof. Baeta Neves"

IMPROVING SPECIES DISTRIBUTION MODELS TO DESCRIBE STEPPE BIRD OCCURRENCE PATTERNS AND HABITAT SELECTION IN SOUTHERN PORTUGAL

by

Pedro Jorge Paixão Leitão, Lic., M.Sc.

Thesis submitted for the degree of Doctor of Philosophy

December 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS

SCHOOL OF CIVIL ENGINEERING AND THE ENVIRONMENT

Doctor of Philosophy

IMPROVING SPECIES DISTRIBUTION MODELS TO DESCRIBE STEPPE BIRD OCCURRENCE PATTERNS AND HABITAT SELECTION IN SOUTHERN PORTUGAL

by Pedro Jorge Paixão Leitão

The birds of the steppe environments face a number of different threats relating to habitat degradation (such as agricultural intensification, land abandonment or afforestation), and the vast majority of species have unfavourable conservation status. Conservation measures require an understanding of species habitat preferences and their occurrence patterns and must be applied at the relevant spatial scales. This study, investigated the habitat selection and resulting distributions of the steppe bird community in southern Portugal, one of its strongholds. To this aim, it made use of large and balanced species location datasets and high quality environmental descriptive data (with a strong emphasis in remote sensing data), collected at two different spatial scales. It applied advanced processing techniques for information extraction, and it optimised the use of species distribution models through a robust methodological framework. Species responses and predictions were derived at the two scales, and a subsequent hierarchical approach was implemented for multi-scale model integration. This procedure resulted in new findings about the ecology of some of the species, particular spatial scale effects were identified and the distributions of the species were further described within the studied area. It also demonstrated that careful planning and the deep understanding of the data analysis methodologies used can result in significant advances in the scientific knowledge. Furthermore, it is recommended that the added-knowledge resulting from this study is incorporated into existing management practices, for an efficient conservation of the species and their habitats.

RESUMO

MELHORAMENTO DE MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES PARA A DESCRIÇÃO DOS PADRÕES DE OCURRÊNCIA E SELECÇÃO DE HABITAT DE AVES ESTEPÁRIAS NO SUL DE PORTUGAL

As aves dos ambientes estepários enfrentam uma série de diferentes ameaças reacionadas com o degradamento dos seus habitats (tais como a intensificação agrícola, o abandono das terras ou a florestação), e a grande maioria das espécies possui um estatuto de conservação desfavorável. Medidas de conservação requerem o conhecimento dos padrões de preferência de habitat e ocurrência e necessitam de ser aplicados nas escalas adequadas. Este estudo investigou a selecção de habitat e resultantes distribiuções da comunidade de aves estepárias no Sul de Portugal, que constitui um dos seus redutos. Para este fim, foram utilizadas bases-de-dados de localização das espécies de grande dimensão e equilibradas, e dados descriptores do ambiente de elevada qualidade (com um grande ênfase em dados de detecção remota), coleccionados a duas escalas espaciais diferentes. Técnicas avançadas de processamento de dados foram aplicadas para extracção de informação, e modelos de distribuição de espécies foram optimisados através de um enquadramento metodológico robusto. As respostas das espécies e suas predições de distribuição foram derivadas às duas escalas e subsequentemente foi implementada uma abordagem hierárquica de integração multi-escala dos modelos. Este procedimento resultou em novas descobertas acerca da ecologia de algumas das espécies, efeitos de escala particulares foram identificados, e as distribuições das espécies foram descritas detalhadamente dentro da área estudada. Este estudo demonstrou também que o planeamento cuidado, juntamente com a profunda compreensão das metodologias de análise de dados utilizadas podem resultar em avanços significativos no conhecimento científico. Ainda, é recomendado que o conhecimento adicionado resultante deste estudo seja incorporado em medidas de gestão existentes, para uma eficiente conservação das espécies e seus habitats.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

There are many people who helped me in many ways before and during the time of the work here presented. In that time I was also supported by several institutions, at various different stages of this work. Without the help and support from all of them, the completion of this thesis would not have been possible, for which I am deeply grateful. Nevertheless, I would like to refer some names as an expression of my special gratitude.

Firstly, I thank my two supervisors, Patrick Osborne at the University of Southampton and Francisco Moreira at the Centro de Ecologia Aplicada "Prof. Baeta Neves", for the guidance, support, motivation and inspiration given on all stages of this work. I also thank them for having believed in my potential from the first moment, having helped me to conceive and prepare the original proposal of this project. They always supported my development and learning with great interest and generosity. The complementary knowledge of both provided the best guidance through this interdisciplinary work. It has been a great privilege to enjoy their supervision, for which I will always be grateful.

This work was funded by a doctoral grant from the FCT (SFRH/BD/12569/2003), having benefitted from additional support from the British Council's Treaty of Windsor Programme / Acções Integradas Luso-Britânicas.

I would like to thank the people who were influential in my early research career, who first taught me the principles of Ecology, Remote Sensing and research work in general. In particular I thank, Prof. Francisco Andrade, Manuel João Pinto, Adelaide Ferreira, Prof. Fernando Catarino and Prof. Jorge Palmeirim.

At the early stages of this project, several periods were spent at the School of Biological and Environmental Sciences of the University of Stirling. I would like

to thank all my colleagues at Stirling, particularly Gilbert Stevenson, Joanne (Thomas) McKenzie and Nicola Schiavottiello, for their friendship and support. Since these times I have also enjoyed the friendship and great inspiration from Dan Chamberlain and Andy Wilson from the BTO Scotland, and Alan Lauder from the RSPB Scotland, to whom I thank.

During the times spent in Southampton, I enjoyed the support from all my colleagues at the Centre for Environmental Sciences, particularly from Colm Bowe, Toru Tsuzaki, Zoe Dunsiger, Moupia Rahman and Simon Bray. The good friendship from Nick, Pete, Salomé, Clive, Andriy, Cecio, Stefano, Lara, Olai and so many others always made my stays away from home much more pleasant. Also my friends in London, particularly Mark, Ouriço, Figura and Jorge, who so many times hosted me and this way facilitated some well-needed weekend breaks. To all of them I am very grateful!

I would like to thank Prof. Francisco Castro Rego, for having hosted me at the Centro de Ecologia Aplicada. Thanks are also due to my colleagues at the Centro, particularly Rui Morgado, Pedro Rocha, Ana Delgado, Susana Dias and Mariana Carvalho for their friendship, support and motivation.

The final phase of this project was spent at the Geomatik Abteilung of the Geographisches Institut, at the Humboldt Universität zu Berlin. I am very thankful to Prof. Patrick Hostert for having hosted me and for some interesting theoretical discussions. My thanks go as well to all my *Geomatiker* colleagues, in particular to Tobias Kümmerle, Andreas Rabe, Sven Lautenbach, Veronique St. Louis, Ruth Sonnenschein and Sören Haubrock for useful knowledge and experience exchanges.

The data collection benefitted from support at many different levels. The first field season (spring of 2004) had the support on logistics and accommodation by SPEA, LPN and Perímetro Florestal da Contenda. The subsequent field

campaigns (springs of 2005 and 2006) were based at the Vale Gonçalinho field station of the LPN Castro Verde. In Castro Verde, I also enjoyed the support and motivation from my fellow colleagues, in particular Rui Morgado, Pedro Rocha, Ana Delgado, Inês Henriques, Inês Moreira, Ana Teresa Marques, Inês Catry, Dave Hill, Rita Alcazar and Luís Reino, and the staff at the field station, Ivone, Fernando, Cátia and Rui. The STEPPEBIRD campaign, held in 2006, was funded by the EUFAR and supported by NERC ARSF, NERC GEF, NERC FSF, LPN, SPEA, PNVG and the ULHT. I also wish to thank all my colleagues who participated in this survey. Additional species locational data were provided by the research projects EDIA/PMo5.4, PRAXIS/C/AGR/11062/1998 and PRAXIS/C/AGR/11063/1998. Environmental descriptive data were provided by the EEA, IEP, the Agência Portuguesa do Ambiente / Atlas do Ambiente Digital and the research project LIFE2002/NAT/P8481.

The data processing and analysis would have not been possible without the help of several people. Prof. Ted Milton (School of Geography, University of Southampton) provided great help during the pre-processing of the airborne data, including access to the NCAVEO lab. Also the help of Dr. Bill Mockridge (Azimuth Systems UK) was critical for the pre-processing of the LiDAR data. The implementation of the SVM image classifications was only possible with the help of Andreas Rabe. Finally, the modified R 'mda' codes used for running the MARS models were kindly provided by Jane Elith (School of Botany, University of Melbourne). To all of them my greatest gratitude!

During the writing-up phase, my thanks go to Maria João Santos (Center for Spatial Technologies and Remote Sensing, University of California Davis) for useful comments in parts of the manuscript, and to Prof. Giles Foody (School of Geography, University of Nottingham) for helpful bibliographic support.

During this work I also benefitted from interesting discussions, during several scientific meetings and workshops, which undoubtedly contributed to my knowledge and understanding of the research methods here applied. Particularly, I would like to thank Jan Elith, Boris Schröder (Institut für Geoökologie, Universität Potsdam), Carsten Dormann (Helmholz - Zentrum für Umweltforschung) and my fellow colleagues from the EBCC Spatial Modelling Work Group, Henk Sirdsema, Lluis Brotons, Richard Gregory, Stewart Newson, Marc Kéry, Frédéric Jiguet and Lechoslav Kuczynski.

I thank my ascending family, in particular my mother and sister Luisa, for always taking me as I am, and in many ways having shaped the person I am today. I also thank my extended family, constituted of my close friends and fellow capoeiristas, for so many shared moments and for having given me so many life lessons - how useful they have been all this time!!

This already very extensive list, however, cannot be complete without acknowledging the most important people. I deeply thank Uli, my beloved one, for always believing in me and giving me so much support. Without her endless love and motivation it would have never been possible to fulfil the tremendous task this work constituted. Instead, she gave some sense in accepting all the challenges ahead. She bared the weight of many long stays away and abroad, of many long nights by the computer, of too many occasions of work-related stress. I will always thank her for her great generosity. Most importantly, with her I learned the meaning of true love, from which we received the greatest gift of all, our beautiful daughter Tara Lia. To them two I dedicate this thesis.

ACRONYMS AND ABBREVIATIONS

AIC	Akaike's Information Criterion
ALTM	Airborne Laser Terrain Mapper
ANN	Artificial Neural Networks
ANOVA	ANalysis Of VAriance
app.	approximately
ARSF	Airborne Research & Survey Facility
ASAR	Advanced Synthetic Aperture Radar
ASCII	American Standard Code for Information Interchange
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATSR	Along-Track Scanning Radiometer
AUC	ROC Area Under the Curve
AUCcv	Cross-validated ROC Area Under the Curve
AVHRR	Advanced Very High Resolution Radiometer
Azgcorr	Azimuth Systems' Geometric correction software
BD	Bolsa de Doutoramento (*Doctoral Grant*)
BIOCLIM	Bioclimatic Modelling, by Busby (1991)
BRT	Boosted Regression Trees
BTO	British Trust for Ornithology
C	SVM regularisation parameter C
CAE	Censo de Aves Estepárias (*Steppe Bird Survey*)
CART	Classification And Regression Trees
CASI	Compact Airborne Spectographic Imager

ca.	circa (*approximately*)
CCA	Canonical Correspondence Analysis
CHRIS	Compact High Resolution Imaging Spectrometer
CI	Confidence Interval
C-SVM	SVM formulation, trained through the definition of the regularization parameter C
DEM	Digital Elevation Model
DGPS	Differential Global Positioning System
DIP	Digital Image Processing
DOMAIN	DOMAIN modelling procedure, by Carpenter *et al.* (1993)
DOS	Dark Object Subtraction
DSM	Digital Surface Model
DTM	Digital Terrain Model
Δ	Delta, symbol to denote difference
E	East
EBCC	European Bird Census Council
EC	European Commission
EDIA	Empresa de Desenvolvimento e Infra-estruturas do Alqueva (*Alqueva's Development and Infra-structures Company*)
EEA	European Environment Agency
EEC	European Economic Community
e.g.	exempli gratia (*for example*)
ENFA	Environmental Niche Factor Analysis, by Hirzel *et al.* (2002)
EnMAP	Environmental Mapping and Analysis Program
ENVISAT	Environmental Satellite

EO	Earth Observing
ERS	European Remote sensing Satellite
ESRI	Environmental Systems Research Institute
et al.	et alli (*and others*)
etc.	et cetera (*and so forth*)
ETM+	Enhanced Thematic Mapper +
EU	European Union
EUFAR	EUropean Fleet for Airborne Research
FCT	Fundação para a Ciência e a Tecnologia (*Foundation for Science and Technology*)
FPH	First Pulse Height
FPI	First Pulse Intensity
ft	foot / feet, non-SI unit of length (1 ft = 0.30480 m)
GAM	Generalised Additive Model
GARP	Genetic Algorithm for Rule-set Production
GCP	Ground Control Point
GCV	Generalized Cross-Validation
GEF	Geophysical Equipment Facility
GEON	Geosciences Network
GIS	Geographical Information System
GLM	Generalised Linear Model
GOES	Geostationary Operational Environmental Satellite
GPS	Global Positioning System
GWR	Geographically Weighted Regression
HRG	High Resolution Geometrical

HRV	Haute Resolution Visible (*High Resolution Visible*)
HRVIR	High Resolution Visible and InfraRed
IBA	Important Bird Area
ICBP	International Council for Bird Preservation
i.e.	id est (*that is*)
IEP	Instituto de Estradas de Portugal (*Portuguese Roads Institute*)
IFOV	Instantaneous Field-Of-View
IGeoE	Instituto Geográfico do Exército (*Portuguese Army Geographical Institute*)
IGP	Instituto Geográfico Português (*Portuguese Geographical Institute*)
IRS	Indian Remote sensing Satellite
ISODATA	Iterative Self-Organizing Data Analysis Technique
IT	Information Theoretic
Itres	Innovative imaging Technology and leading edge Research to meet the needs of customers with emphasis on scientific Excellence and Service
JNCC	Joint Nature Conservation Committee
Jr.	Junior
km	kilometre, SI unit of length (1 km = 1000 m)
LC	Land Cover
LIBSVM	A Library for Support Vector Machines, by Chang & Lin (2001)
LiDAR	Light Detection And Ranging
LISS	Linear Imaging and Self Scanning
LOF	Lack-Of-Fit
LPH	Last Pulse Heigth

LPI	Last Pulse Intensity
LPN	Liga para a Protecção da Natureza (*Nature Protection League*)
LSMM	Linear Spectral Mixture Model
m	metre, SI basic unit of length
MARS	Multivariate Adaptive Regression Splines
MaxEnt	Maximum Entropy, by (Phillips *et al.* 2006)
mda	Mixture Discriminant Analysis package for R, by Hastie & Tibshirani (1996)
MERIS	MEdium Resolution Imaging Spectrometer
METEOSAT	Meteorological Satellite
MLC	Maximum Likelihood Classifier
MODIS	Moderate Resolution Imaging Spectroradiometer
MRT	Multivariate Regression Trees
MS	Multi-scale
MVC	Maximum Value Composite
MVIRI	Meteosat Visible and Infrared Imager
n	Sample size
N	North
NASA	National Aeronautics and Space Administration
NCAVEO	Network for Calibration And Validation in Earth Observation
NDMI	Normalized Difference Moisture Index
NDSI	Normalized Difference Soil Index
NDSVI	Normalized Difference Senescent Vegetation Index
NDVI	Normalized Difference Vegetation Index
NE	Northeast

NERC	Natural Environment Research Council
NHC	Natural History Collections
NOAA	National Oceanic and Atmospheric Administration
NSMI	Normalized Soil Moisture Index
n.s.	non-significant
OCA	Overall Classification Accuracy
OSAVI	Optimized Soil-Adjusted Vegetation Index
p	Statistical test p value of significance
PC	Principal Component
PCA	Principal Component Analysis
PDF	Probability Density Function
pH	Power of hydrogen, measure of acidity
PNVG	Parque Natural do Vale do Guadiana (*Guadiana Valley Natural Park*)
PROBA	PRoject for On Board Autonomy
PSF	Point Spread Function
r	Pearson r correlation coefficient
R	R Statistical analysis software, by R Development Core Team (2008)
RADARSAT	Radar Satellite
RDA	Redundancy Analysis
RF	Random Forests
ROC	Receiver Operating Characteristic
rs	Spearman rho rank correlation coefficient
RS	Remote Sensing
RSPB	Royal Society for the Protection of Birds

S	South
SAR	Synthetic Aperture Radar
SDM	Species Distribution Model(ling)
SE	Standard Error of the mean
SEO	Sociedad Española de Ornitología (*Spanish Ornithological Society*)
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SFRH	Serviço de Formação dos Recursos Humanos (*Service for Human Resources Training*)
SI	Système International d'Unités (*International System of Units*)
SMA	Spectral Mixture Analysis
SOM	Self Organizing Map
SPA	Special Protection Area for birds
SPEA	Sociedade Portuguesa para o Estudo das Aves (*Portuguese Society for the Study of Birds*)
SPOT	Satellite Probatoire d'Observation de la Terre (*Earth Observation Probe Satellite*)
spp.	species, indicating several species of the respective genus
SRTM	Shuttle Radar Topography Mission
StepGAM	Stepwise Generalized Additive Model
SVM	Support Vector Machines
TC	Tasseled Cap or the Kauth-Thomas transformation (1976)
TM	Thematic Mapper
τ	Kendall's tau rank correlation coefficient
U	Mann-Whitney U rank test value
ULHT	Universidade Lusófona de Humanidades e Tecnologias (*Lusófona University of Humanities and Technologies*)
UK	United Kingdom

UTM	Universal Transverse Mercator
USA	United States of America
VCM	Varying Coefficient Modelling
VGT	VEGETATION, sensor onboard the SPOT satellite
VHM	Vegetation Height Model
WGS84	World Geodetic System 1984

CHAPTERS

1. *Introduction*

In Europe, low-intensity farming systems have the highest bird diversity of all agricultural systems (Beaufoy *et al.* 1994; Bignal & McCracken 1996; Tucker 1997). Being dominated by cereal steppe landscapes (pseudo-steppes), the south of the Iberian Peninsula, holds significant numbers of several threatened bird species, such as the Great Bustard (*Otis tarda*), the Little Bustard (*Tetrax tetrax*), the Black-bellied Sandgrouse (*Pterocles orientalis*) and the Lesser Kestrel (*Falco naumanni*) (Tucker & Heath 1994; Suárez *et al.* 1997; BirdLife International 2004; Burfield 2005; Santos & Suárez 2005). In Portugal, these flat and open landscapes are created by the extensive cultivation of cereals on a rotational basis. The result is a mosaic of cereal fields, stubbles, ploughed and fallow land, the latter being used as sheep pastures (Suárez *et al.* 1997; Moreira 1999). In the last decades the land use of these steppes has been changing due to agricultural intensification (through irrigation) as well as land abandonment and afforestation (Tucker & Heath 1994; Burfield 2005; Santos & Suárez 2005). This habitat degradation and loss has a direct impact on the bird populations as has been described by several authors (Baldock 1991; Tucker 1991; Tucker & Heath 1994; Suárez *et al.* 1997; Burfield 2005; Santos & Suárez 2005).

In 1992, the European Union introduced the principle of maintaining these extensive farming systems under the Agri-Environment Programme (EU Regulation 2078/92). This incorporates compensation for farmers for keeping agricultural practices that allow the conservation of threatened species (Robson 1997; De la Concha 2005), but their effectiveness is open to question (Kleijn & Sutherland 2003; Berendse *et al.* 2004; Kleijn *et al.* 2004; Whittingham 2007). The new European Union strategic guidelines for Rural Development for the period 2007-2013 (2006/144/EC) incorporate these agri-environment schemes within the scope of two main EU priorities, the Sustainable Development Strategy and the Gothenburg commitment to halt the loss of biodiversity by 2010. These reforms have generally been taken both with optimism and concern, as they will fundamentally change the development and management of the agricultural semi-natural landscapes that constitute the pseudo-steppes of Iberia (De la Concha 2005; Oñate 2005). In order to apply adequate management schemes for the conservation of steppe birds and habitats, it is necessary to understand the way these species use the environment. Hence there is a need for scientific studies aimed at a better understanding of these agricultural landscapes and the effects of human use on bird's habitats.

Recently, many studies have focused on the steppe environment and its avifauna (Tellería *et al.* 1988; Martinez & Purroy 1993; Beaufoy *et al.* 1994; Martínez 1994; Suárez *et al.* 1997; Moreira 1999; Delgado & Moreira 2000; Lane *et al.* 2001; Suárez-Seoane *et al.* 2002a; Wolff *et al.* 2002; Brotons *et al.* 2004a; Franco & Sutherland 2004; Silva *et al.* 2004; Franco *et al.* 2005; Moreira *et al.* 2005; Pinto *et al.* 2005; Morales *et al.* 2006; Osborne & Suárez-Seoane *2007;* Silva *et al.* 2007; Traba *et al.* 2007), denoting a growing acknowledgement of its importance for biodiversity conservation. Indeed, 83% of the steppe birds in Europe have an unfavourable conservation status, which is twice the overall figure for all European birds (BirdLife International 2004; Burfield 2005).

None of the existing studies, however, has focused on the issue of spatial scale and its effects on habitat use and selection, and the resulting occurrence patterns of steppe birds at different scales. Scale is a key issue in all the spatial and environmental sciences. Ecological processes and physical characteristics possess an inherent scale at which they occur (Turner *et al.* 1989; Wiens 1989; Levin 1992; Whittaker *et al.* 2001; Blackburn & Gaston 2002; Boyce 2006). Also, the pattern detected in an ecological mosaic is a function of scale. Thus, scale is important in several aspects of the study of landscapes, from factors affecting single organisms to continental plate tectonics (Forman & Godron 1986; Carlile *et al.* 1989; Foody & Curran 1994). Additionally, human activities are increasingly affecting patterns and processes at many different scales while both animals and humans generally perceive and respond to only a fraction of the multi-scale heterogeneity present in natural systems (Forman 1995; Farina 1998).

The major importance of scale has led many authors to conclude that ecological processes and patterns should be studied across scales by using a multi-scale (MS) approach (Gardner & Turner 1991; King 1991; Gordon & Dennis 1996;

Farina 1998; Gering *et al.* 2003; Meyer & Thuiller 2006). This is an obvious research gap in studies of agricultural steppe birds. A better understanding of the way species respond to the environment across different scales has a profound consequence for conservation as it allows the development of better management measures. Such knowledge would permit a better evaluation of the impacts of agricultural uses on the steppe environment and a subsequent optimisation of the EU subsidies aimed at nature conservation. Conservation needs habitat management prescriptions for threatened steppe birds on agricultural land that incorporate scale effects (see e.g. Poiani *et al.* 2000). Scale can refer to both spatial and temporal scale and can be expressed by grain and extent. For the spatial scale, grain refers to the spatial resolution, i.e. the pixel size, and extent refers to the size of the study area (Turner *et al.* 1989). From an agricultural management perspective, spatial scale can be reflected in terms of practices implemented at the patch/parcel level up to the landscape level.

Geographical Information Systems (GIS) are capable of efficiently storing and managing ecosystems data for large areas, as well as translating data between multiple scales (Stow 1993; Longley 1998). Remote sensing (RS) data analyses explore the spectral characteristics of the Earth's surface through complex mathematical algorithms and procedures. With different sources of imagery, it is possible to infer resource patterns and habitats at multiple scales through time (Quattrochi & Pelletier 1991; Campbell 1996; Jensen 1996; Curran *et al.* 1998). Thus, GIS integrating remotely sensed data provides ecologists with a powerful tool for carrying out numerical modelling of spatial ecosystem processes (Palmeirim 1988; Stow 1993; Eastman *et al.* 1995; Johnston 1998; Osborne *et al.* 2001; Osborne 2005). Many studies have made use of these techniques for addressing ecological questions (Palmeirim 1988; Johnston 1989; Pienkowski *et al.* 1989; Griffiths *et al.* 1993; Tucker *et al.* 1997; Brito *et al.* 1999; Osborne *et al.* 2001; Leitão *et al.* 2002; Suárez-Seoane *et al.* 2002a; Suárez-Seoane *et al.* 2004; Fuller *et al.* 2005; Osborne & Suárez-Seoane *2007*). Several recent papers review the use of RS data for biodiversity mapping (Nagendra 2001; McDermid *et al.* 2005; Leyequien *et al.* 2007; Gillespie *et al.* 2008). Also, Gottschalk *et al.* (2005)

provide a good review of bird habitat and distribution modelling studies using satellite imagery over the last 30 years.

This study builds on this background by examining the patterns of habitat selection and occurrence by steppe birds during the breeding season at two different spatial scales, within a common methodological framework.

1.1. Methodological concept

This study was performed at two different scales: regional and landscape. At each particular scale there are four main work components: habitat characterisation; bird population characterisation; GIS integration and modelling; and model interpretation. Habitats were characterised by RS imagery, map data, and field measurements. The bird populations were characterised by bird occurrence (presence / absence) data collected in the field, during the breeding season. A specific methodology for data collection was used at each scale. All the data were integrated in a GIS, to be used in empirical habitat models, at both spatial scales.

Habitat models based on empirical species data and habitat/environmental descriptors quantify the associations between the species and their habitats, and in this way describe their habitat preferences and requirements. This can be particularly relevant for understanding a species' ecology and its likely response to environmental change. When habitat data are available for a certain region, these models can then be used to predict a species' abundance and potential distribution by mapping the areas of suitable habitat. In recent times, the use of habitat models for the spatial prediction of species' distributions has increased greatly, and they are now considered a significant tool in conservation planning and wildlife management (Buckland & Elston 1993; Guisan & Zimmermann 2000; Austin 2002; Guisan & Thuiller 2005). In fact, most of the recent developments in habitat models have been published in the species distribution modelling literature, and thus are hereafter referred to as Species Distribution Models (SDMs).

The modelling methodologies depend partly on the available species' data, which can typically be presence-only (derived from field-based observations, herbarium data, etc.), presence-absence (from field surveys) or abundance (count) data. Different statistical approaches include: a) generalised regressions, such as Generalised Linear Models (GLM; McCullagh & Nelder 1989), Generalised Additive Models (GAM; Hastie & Tibshirani 1990; Guisan *et al.* 2002) and Multivariate Adaptive Regression Splines (MARS; Friedman 1991; Leathwick *et al.* 2006); b) classification techniques, such as Multivariate Regression Trees (MRT; De'ath 2002) and Classification And Regression Trees (CART; Vayssières *et al.* 2000); c) environmental envelope models, like BIOCLIM (Busby 1991) or DOMAIN (Carpenter *et al.* 1993); d) ordination techniques, such as Canonical Correspondence Analysis (CCA; ter Braak 1986; Guisan *et al.* 1999) and Environmental Niche Factor Analysis (ENFA; Hirzel *et al.* 2002); e) Bayesian approaches, i.e., based on Bayesian statistics (Bayes & Price 1763; Tucker *et al.* 1997; Termansen *et al.* 2006); and f) machine learning techniques, like Genetic Algorithm for Rule-set Production (GARP; Stockwell & Peters 1999), Artificial Neural Networks (ANN; McCulloch & Pitts 1943; Tan & Smeins 1996), Support Vector Machines (SVM; Cortes & Vapnik 1995; Drake *et al.* 2006), Random Forests (RF; Breiman 2001), Boosted Regression Trees (BRT; De'ath 2002; Elith *et al.* 2008) or Maximum Entropy Modelling (MaxEnt; Phillips *et al.* 2006; Phillips & Dudík 2008).

Several recent studies review and compare different techniques, their functioning, applicability and performance (Guisan & Zimmermann 2000; Moisen & Frescino 2002; Brotons *et al.* 2004b; Segurado & Araújo 2004; Elith *et al.* 2005; Elith *et al.* 2006; Lawler *et al.* 2006), which is not the aim of the present study. Instead, a single approach is used to fit the species occurrence data to the environmental predictors at both scales of study. This eliminates one source of uncertainty when comparing patterns across scales. To this end MARS models were used, which are capable of fitting complex functions and have been described as a robust method, capable of achieving high model performances with fast computation speeds (Moisen & Frescino 2002; Leathwick *et al.* 2006). Additionally, like other regression model approaches, MARS allows the generation of partial regression plots, and thus facilitates the inference of the species responses to the environmental descriptors (predictor variables).

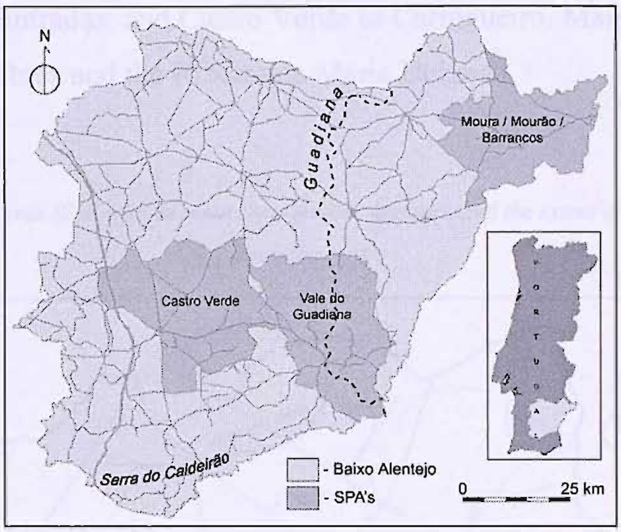
In order to incorporate the effects of (spatial) scale in the investigation of the species habitat preferences, SDMs were built at both regional and landscape scales. Finally, MS models were built in a hierarchical manner, by incorporating the regional scale model predictions as predictor variables in the landscape scale models (see Chapter 5).

1.2. Study area

This study moves across scales by changing both grain and extent of analysis, the latter implying a change in the extent of the study area. Thus, the study area at the regional scale is the Baixo Alentejo region in southern Portugal, which covers an area of approximately 8500 km^2 (Figure 1.1), and is characterised by dry conditions (Mean Annual Precipitation: 400-800mm) and high temperatures (Average Maximum Temperature: 13-30°C). It reflects the typical diversity and structure of semi-natural Mediterranean landscapes, with features such as cereal steppe grasslands (or pseudo-steppes), open woodland, shrublands, olive-groves and vineyards (Neves 1998).

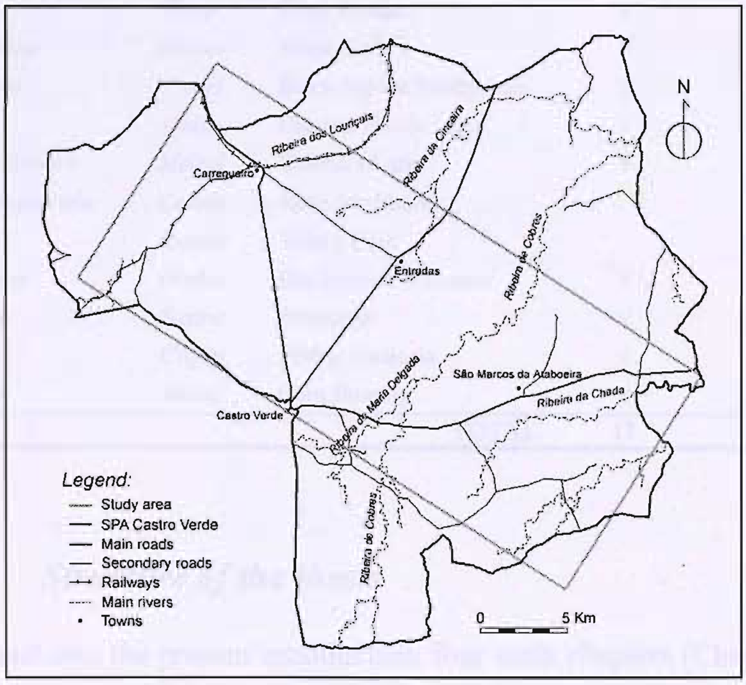
The main topographic features consist of the valley of the Guadiana River and the Serra do Caldeirão hill chain on the south-western edge. Otherwise, most of the area is either flat or slightly undulating. It has a low human population density, and is served by a small road network, with one highway crossing the western edge of the area. The region is of national importance for steppe birds and includes three Special Protection Areas (SPAs) for birds (EU directive 79/409/EEC) with importance for breeding steppe birds (Costa *et al.* 2003).

Figure 1.1 - The Baixo Alentejo region in southern Portugal (light grey, inset), with its main features: the Guadiana River (black dashed line) and its valley; the Serra do Caldeirão hill chain; road network (grey lines); and SPAs (in darker grey)



The landscape scale study focused on an area of approximately 47500 ha covering most of the open (steppe) area within the Castro Verde SPA (Figure 1.2). The Castro Verde SPA landscape is a rolling plain (100-300m) of about 80,000 ha, with a Mediterranean climate, including hot summers (30-35°C on average in July), fairly cold winters (averaging 5-8°C in January) and over 75% of the annual rainfall (500–600 mm) concentrated in the October-March months (Moreira *et al.* 2005). It is classified under the Natura2000 sites network, and is designated as an Important Bird Area (IBA) (Grimmet & Jones 1989; Costa *et al.* 2003). It constitutes the main area of cereal steppes (pseudo-steppe habitats) in the country and is of national and international importance for populations of several threatened steppe bird species, such as the Great Bustard, Little Bustard, Black-bellied Sandgrouse, and Lesser Kestrel, among others (Moreira 1999; Costa *et al.* 2003; Pinto *et al.* 2005; Moreira *et al.* 2007) (see Appendix A.1). A great proportion of the SPA (ca. 55,000 ha) is covered with cereal steppes, but also includes some areas of shrublands (mostly *Cistus* spp.), mainly associated with river valleys and in the southeastern part of the region, interspersed with old fallows resulting from agricultural abandonment and scrub encroachment (Delgado & Moreira 2000; Moreira *et al.* 2007) (see Appendix A.1). Less frequent land uses that can be found in the area are some recent afforestations of eucalyptus (*Eucalyptus* spp.), umbrella pines (*Pinus pinea*) and holm oak (*Quercus rotundifolia*) (Moreira *et al.* 2005; Moreira *et al.* 2007) (see Appendix A.1). Three main roads cross the area: Castro Verde to São Marcos da Ataboeira; Castro Verde to Entradas; and Castro Verde to Carregueiro. Main rivers include the Ribeira de Cobres and the Ribeira de Maria Delgada.

Figure 1.2 - Castro Verde SPA, with its main cartographic features, and the extent of the study area at the landscape scale



1.3. Target species

The study focused on the steppe bird community, which uses the agricultural landscapes of the Baixo Alentejo and particularly the pseudo-steppes of the Castro Verde SPA. In total, 15 species were considered, as listed in Table 1.1. The Crested and Thekla larks (*Galerida cristata* and *Galerida theklae*) were, however, categorised to the genus level (*Galerida* spp.) due to difficulties in the reliable identification of all individuals of these two species in the field (Moreira *et al.* 2007) (see Appendix A.1). In the case of some of the species, the lack of presence data points at a particular scale dictated its exclusion from the respective study. The performance of the regional scale models defined the list of species to be included in the MS analysis (see Chapter 5).

Table 1.1 - Species considered at each scale of study

Scientific name	**Acronym**	**Common name**	**Regional**	**Landscape**	**MS**
Circus pygargus	*Cirpyg*	Montagu's Harrier	√	√	√
Alectoris rufa	*Aleruf*	Red-legged Partridge	√	√	
Coturnix coturnix	*Cotcot*	Quail	√		
Tetrax tetrax	*Tettet*	Little Bustard	√	√	√
Otis tarda	*Otitar*	Great Bustard	√	√	√
Burhinus oedicnemus	*Buroed*	Stone Curlew	√	√	
Pterocles orientalis	*Pteori*	Black-bellied Sandgrouse	√		
Galerida spp.	*Galsp*	Crested/Thekla Lark	√	√	
Melanocorypha calandra	*Melcal*	Calandra Lark	√	√	√
Calandrella brachydactyla	*Calbra*	Short-toed Lark		√	
Anthus campestris	*Antcam*	Tawny Pipit		√	
Oenanthe hispanica	*Oenhis*	Black-eared Wheatear	√	√	
Saxicola torquata	*Saxtor*	Stonechat	√	√	
Cisticola juncidis	*Cisjun*	Zitting Cisticola	√	√	√
Miliaria calandra	*Milcal*	Corn Bunting	√	√	√
		TOTAL	13	13	6

1.4. Structure of the thesis

This thesis includes the present introduction, four main chapters (Chapters 2 - 5) and a final synthesis (Chapter 6). Chapters 2 to 4 refer to specific problems relating to data analysis, either the functioning of the SDMs or the processing of RS data for feature extraction. Chapter 2 uses the regional scale data to assess the influence of data sampling bias on the performance of SDMs and on the respective predicted distributions patterns. This chapter is written in a manuscript form as it constitutes a draft for submission to publication in a peer-reviewed scientific journal. Chapter 3 explores (at the landscape scale) different possible Digital Image Processing (DIP) approaches for extracting SDM predictor variables from Landsat TM data, and their influence on model performance and interpretation, while also reviewing some of the major sources of RS data (to be used as environmental descriptors). Chapter 4 describes the STEPPEBIRD airborne campaign and the respective processing of high-resolution laser altimetry (or LiDAR - Light Detection And Ranging) data for feature extraction in the Castro Verde study area. Chapter 5 includes the final SDMs and respective interpretation of the species distribution patterns and habitat preferences at the two spatial scales considered, as well as the MS model integration for some of the studied species. The last chapter (Chapter 6) interprets the findings of the previous chapters in a unified context and discusses some possible directions for further research.

Additionally, two appendices are included, which refer to work done in parallel to that within this thesis, and further explore some of the aspects covered in this work. Appendix A.1 includes the results of the "CAE - Censo de Aves Estepárias", a survey of the steppe birds of Castro Verde which occurred during the Spring 2006 and which provided data to the landscape scale study (see Chapter 5). Appendix A.2 uses data collected within the regional scale study surveys and explores the influence of species and habitat positional errors on SDMs, their performance and interpretation.

1.5. Publication overview

Several parts of this work have been presented in International Conferences, Symposia and Workshops, as follows:

Pedro J. Leitão, Patrick E. Osborne & Francisco Moreira. *Habitat-based distribution modelling of agricultural steppe birds in South Portugal – a multi-scale approach.* 16th European Bird Census Council "Bird Numbers 2004" International Conference, held in Kayseri, Turkey, from the 6th to the 11th of September 2004 – Oral presentation.

Pedro J. Leitão, Patrick E. Osborne & Francisco Moreira. *Predicting steppe-land bird distributions in Baixo Alentejo, Portugal.* International Symposium on Ecology and Conservation of Steppe-land Birds, held in Lleida, Spain, from the 3rd to the 7th December 2004 – Poster presentation.

Pedro J. Leitão, Patrick E. Osborne & Francisco Moreira. *Multi-scale habitat selection and distribution modelling of steppe birds in Baixo Alentejo, Portugal.* International Workshop "Predictive modelling of Species distribution - New Tools for the XXI Century", within the series of "Current Trends in Environment Workshops", held in Baeza, Spain, from the 2nd to the 4th of November 2005 – Poster presentation.

Pedro J. Leitão, Patrick E. Osborne & Francisco Moreira. *The use of large-scale remote sensing and map data to determine steppe-land bird distributions in Baixo Alentejo, Portugal.* 24th International Ornithological Congress, held in Hamburg, Germany, from 13th to the 19th of August 2006 – Poster presentation[1].

Patrick E. Osborne & **Pedro J. Leitão**. *Issues associated with the use of remote sensing data in predictive models of species distributions.* 1st European Congress of Conservation Biology "Diversity for Europe", held in Eger, Hungary, from the 22nd to the 26th of August 2006 – Oral presentation.

Pedro J. Leitão & Patrick E. Osborne. *Effects of misregistered data on species distribution models.* International workshop "Advances in statistical

[1] The abstracts of all congress' presentations were published as a supplement issue of a peer-reviewed scientific journal: *Journal of Ornithology*, 147 (Suppl.1) – see author's declaration at the beginning of this thesis.

modelling of faunal distribution: Global and local applications", held in Giessen, Germany, from the 19th to the 21st of November 2006 – Poster presentation.

Pedro J. Leitão, Patrick E. Osborne & Francisco Moreira. *Remote sensing data for spatial modelling of bird distributions and abundances: potential uses and limitations*. 17th European Bird Census Council "Bird Numbers 2007" International Conference, held in Chiavenna, Italy, from the 17th to the 22nd of April 2007 – Oral presentation.

Pedro J. Leitão, Francisco Moreira, Rui Morgado, Rita Alcazar, Ana Cardoso, Carlos Carrapato, Ana Delgado, Pedro Geraldes, Luís Gordinho, Inês Henriques, Miguel Lecoq, Domingos Leitão, Ana T. Marques, Rui Pedroso, Ivan Prego, Luís Reino, Pedro Rocha, Ricardo Tomé & Patrick E. Osborne. *CAE Castro Verde 2006 - A steppe bird monitoring scheme*. 17th European Bird Census Council "Bird Numbers 2007" International Conference, held in Chiavenna, Italy, from the 17th to the 22nd of April 2007 – Poster presentation.

Pedro J. Leitão, Edward J. Milton, Bill Mockridge, Patrick E. Osborne & Francisco Moreira. *Pre-processing issues affecting the use of CASI and LiDAR data for steppe bird habitat monitoring and management in southern Portugal*. NERC Airborne Research & Survey Facility Workshop within the Annual Conference of the Remote Sensing and Photogrammetry Society, held in Newcastle upon Tyne, from the 11th to the 14th of September 2007 – Oral presentation.

Chapters 2 to 5 of the work here presented are intended to be extended, improved and re-edited with the aim of being submitted for publication in peer-reviewed journals.

Additionally, the papers presented in the appendices section have been submitted for publication in peer-reviewed journals. The first one has been published and the second has been accepted and is currently in press, with the following references:

Moreira, F, **Leitão, P.J.**, Morgado, R., Alcazar, R., Cardoso, A., Carrapato, C., Delgado, A., Geraldes, P., Gordinho, L., Henriques, I., Lecoq, M., Leitão, D., Marques, A.T., Pedroso, R., Prego, I., Reino, L., Rocha, P., Tomé, R. & Osborne, P.E. 2007. Spatial distribution patterns, habitat correlates and population estimates of steppe birds in Castro Verde. *Airo*, 17: 5-30.

Osborne, P.E. & **Leitão, P.J.** *In press.* The effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions.*

2. *Effects of data sampling bias on species distribution models and spatial patterns of biodiversity: a case-study with steppe birds in southern Portugal*

2.1. *Abstract*

Species distribution models provide useful information about species and biodiversity spatial patterns, which form the basis of many ecological applications and management decisions such as the definition of conservation priorities and reserve selection. These models however, are frequently based on existing datasets which have been collected in an unbalanced (biased) manner. In this study we investigated the effects of data sampling bias on model performance, interpretation and predictions. A large steppe bird dataset was collected in southern Portugal, following a carefully designed sampling scheme, and then sub-sampled this dataset with varying degrees of geographical bias and random sampling intensity. Sampling bias was characterised by two measures of sample size and environmental bias. MARS models were run on all datasets, and all the subset models compared with the control in order to assess the effect of the respective bias.

It was found that data sampling bias affected the performance of the models to a varying degree. Environmental bias was generally more influential on the model results than sampling intensity. Model structure, however, was severely affected by sampling bias, raising concern about the ecological interpretations of models run on biased datasets. The resulting output predictions from the models, hence the spatial patterns of species occurrence and biodiversity, were also affected by sampling bias. It is therefore important that special attention is paid to the quality of existing datasets used in distribution models, as well as the sampling design for collection of new data. Also, when modelling with biased datasets, the ecological interpretation of such models should be made with caution and explicit awareness of the existing bias.

2.2. *Introduction*

Understanding spatial patterns of biodiversity is an issue of great concern among ecologists, conservationists and biogeographers, in particular for identifying conservation priorities and optimum nature reserve design (Pressey *et al.* 1993; Ferrier 2002), as well as for monitoring environmental change (Wilson *et al.* 2004).

When detailed species occurrence data are not available, coarse-scale species range maps are commonly used to assess species richness (the most commonly used measure of biodiversity) typically by overlapping range maps of several species (Blackburn & Gaston 1996). An implicit assumption underlying species range maps is that they represent areas of uniform occurrence, which is a source of uncertainty for determining species richness patterns. Range maps generally overestimate species distributions by assigning false presences (errors of commission) within the interior of the species ranges. They therefore lack the local and regional patterns necessary for the interpretation of coarse-scale trends and environmental associations (La Sorte & Hawkins 2007).

As an alternative to using crude range maps, detailed species distributions can be inferred by using empirical predictive models (Guisan & Zimmermann 2000). The use of predictive models has expanded greatly in ecological research, benefiting from advances in statistical techniques, as well as from continued software and hardware developments. These, particularly when coupled with RS data capable of describing environmental conditions in a detailed, systematic and synoptic manner, have been shown to be very useful tools for describing biodiversity patterns (Fuller *et al.* 1998; Kerr & Ostrovsky 2003; Leitão *et al.* 2006). In fact, species distribution models (SDMs) are capable of reducing the frequency of false absences of raw locational datasets, while not incurring the false presences of coarse-scale range maps (Ferrier 2002). However, predictions of species occurrences from distribution models always have a level of error or uncertainty inherent to the modelling process itself. This uncertainty should be acknowledged for appropriate interpretation of the data – particularly important when these models form the basis of a decision-making process (Elith *et al.*

2002). Error in predictions can be the result of many factors, such as small sample sizes, data sampling bias, data prevalence, lack of presence records, mismatched scales in the data, or the failure to incorporate critical habitat variables in the models (Pearce *et al.* 2001; Kadmon *et al.* 2003; Barry & Elith 2006). It has also been noted by some authors that different modelling approaches are likely to generate different spatial patterns, which raises some concerns (Wilson *et al.* 2005). In addition, the ecological characteristics of the modelled species can also have an effect on model accuracy potential (McPherson & Jetz 2007).

In this study I investigate the effects of data sampling bias on SDMs and on the spatial patterns of biodiversity derived from them. Data sampling has been recognised as one of the priority areas for development and research in respect to species distribution modelling (Araújo & Guisan 2006). Existing studies have investigated the role of sample size (Stockwell & Peterson 2002; Wisz *et al.* 2008), sampling survey design (Edwards Jr. *et al.* 2006) and bias (Kadmon *et al.* 2004) on model performance. Others have explored the biased nature of Natural History Collection (NHC) datasets (Graham *et al.* 2004), their effect on model performance (Loiselle *et al.* 2008), and the implications of data sampling bias in ecological research (Martinez & Wool 2006), as well as for defining conservation priorities and reserve design (Reddy & Dávalos 2003; Grand *et al.* 2007). None of these studies has, however, explored the consequences of data sampling on the predicted spatial patterns of biodiversity, which form the base knowledge of many ecological applications and management decisions. I assembled a dataset of steppe bird occurrence/non-occurrence in southern Portugal, based on a carefully designed, intensive and stratified random sampling scheme, which we considered to be complete and unbiased. We then degraded this dataset by sub-sampling to generate realistic biases typically found in species locational datasets. The models were run on all datasets, and model performance and consistency, as well as the predicted maps of species probability of occurrence and respective biodiversity maps, were compared between the complete and biased datasets.

2.3. Methods

• *Study area*

This study was conducted in the Baixo Alentejo region in South Portugal, which covers an area of approximately 8500 km^2 (Figure 1.1) and is characterised by dry conditions (Mean Annual Precipitation: 400-800mm) and high temperatures (Average Maximum Temperature: 13-30ºC). It reflects the typical diversity and structure of semi-natural Mediterranean landscapes, with features such as cereal steppe grasslands, open woodland, shrublands, olive-groves and vineyards (Neves 1998).

The typical cereal pseudo-steppe landscape is a spatio-temporal mosaic of dominant fallow fields with some low-intensity (winter) cereal crops (mostly wheat and oats, but also some barley), which become stubbles after harvest, and are ploughed before seeding. The main topographic features consist of the valley of the Guadiana River and the Serra do Caldeirão hill chain on the south-western edge. Otherwise, most of the area is either flat or slightly undulating. It has a low human population density, and is served by a small road network, with one highway crossing the western edge of the area. The region is of national importance for steppe birds and includes three Special Protection Areas (SPAs) for birds (EU directive 79/409/EEC) with importance for breeding steppe birds (Costa *et al.* 2003). One of these, the Castro Verde SPA, is considered the main steppe area in the country and is of international importance for several steppe bird species (Moreira *et al.* 2007) (see Appendix A.1).

• *Biological data sampling*

The data survey design – an intensive random sample imposed on a systematic geographical stratification – provided a statistical balanced data-base which is representative of the species' frequency of occurrence (natural prevalence) and their distribution patterns in the region, and can be considered close to ideal for distribution modelling (Hirzel & Guisan 2002; Araújo & Guisan 2006). During the spring of 2004 (from the end of March until the middle of June) 560 grid squares of 1x1 km^2 were sampled, equivalent to c. 6% of the total area (Figure 2.1 h). Each square was surveyed once during the early morning or evening (peak activity period of the birds), for a duration of 30 minutes, and the species' occurrence status (presence or absence) determined. Birds were considered present when detected by visual or auditory cues. Twelve species representative of the regional steppe bird community were sampled (Table 2.1). This unbiased dataset, the control, will serve as a reference for comparison with all datasets generated with induced biases.

- *Environmental data*

All environmental layers were computed at a spatial grain of 1x1 km^2 and derived from RS and map data. Following findings from previous studies on steppe birds, data related to vegetation, terrain and human disturbance were used as predictor variables (Osborne *et al.* 2001).

Table 2.1 - List of species studied, respective frequency of occurrence in the area, and model performance on the control dataset (calculated on the training and cross-validated data)

Species	**Freq. of occurrence**	**AUC**	**AUCcv**
Montagu's Harrier	0.188	0.83	0.75
Red-legged Partridge	0.519	0.77	0.66
Quail	0.461	0.80	0.74
Little Bustard	0.285	0.84	0.77
Great Bustard	0.058	0.86	0.78
Stone Curlew	0.136	0.81	0.66
Black-bellied Sandgrouse	0.022	0.94	0.74
Crested/Thekla Lark	0.512	0.73	0.69
Calandra Lark	0.132	0.83	0.76
Black-eared Wheatear	0.055	0.72	0.63
Stonechat	0.440	0.60	0.52
Zitting Cisticola	0.637	0.81	0.73

Vegetation was described by using a temporal series of Normalized Difference Vegetation Index (NDVI) imagery, which allows the characterisation of the vegetation phenology and crop cycle (Hill & Donald 2003). A 12-month series of NDVI data from the SPOT VEGETATION sensor at a spatial resolution of 1 x 1 km was used in the form of 10-day synthesis images (www.spot-vegetation.com/vegetationprogramme/), for the period June 2003 to May 2004. These images, which had been previously atmospherically corrected, were produced by Maximum Value Compositing (MVC) (Maisongrande *et al.* 2004). This method minimises the cloud cover in NDVI imagery and reduces Sun-angle, shadow, aerosol and water-vapour effects, all of which could reduce data reliability (Holben 1986). We excluded all data pixels with reported quality problems, and subsequently used a further MVC procedure to produce a time-series of 12 Monthly-synthesis images. Finally, the time-series was reduced into seven variables, capable of describing the observed phenological events of the vegetation in the area (Table 2.2).

Table 2.2 - Predictor variables used in the species distribution models

Variable	Description
Vegetation	
Summer	Vegetation senescence during the Summer months: *Summer = NDVI (Jun) – NDVI (Sep)*
Winter	Vegetation growth during the Autumn and Winter months: *Winter = NDVI (Mar) – NDVI (Sep)*
Spring	Vegetation senescence during the Spring months: *Spring = NDVI (Jan) – NDVI (Apr)*
Dry	Mean NDVI during the dry months: *Dry = Average [NDVI (Jun : Oct)]*
Wet	Mean NDVI during the wet months: *Wet = Average [NDVI (Jan : Apr)]*
Dec	NDVI value for the month of December: *NDVI (Dec)*
May	NDVI value for the month of May: *NDVI (May)*
Terrain	
Alt	Mean altitude in metres within a 5 x 5 array of 200 x 200 m pixels
Topovl0	Variation in altitude in a 5 x 5 array of 200 x 200 m pixels, where altitude is re-classed to a 10 m vertical resolution. *Topovl0 = (n-1)/(p-1)*, where n = number of different altitude classes in the array, p = number of pixels in the array, i.e. 25
Disturbance	
Urbandist	Distance (in metres) to the nearest pixel containing towns, settlements or constructed structures
Roaddist	Distance (in metres) to the nearest pixel containing roads

Terrain variables were derived from a Digital Terrain Model (DTM) acquired from the Instituto Geográfico Português (IGP), originally at a spatial resolution of 200 x 200 m. From these data, I calculated average altitude (ALT) and topographic variability with a 10-m vertical resolution (TOPOV10) for each grid square (Suárez-Seoane *et al.* 2002a). The disturbance variables considered were distance to towns, urban settlements and constructed structures (URBANDIST) derived from the Corine Land Cover 2000 raster data provided by the European Environment Agency (EEA) and distance to roads (ROADDIST) derived from a vector-based road map provided by the Instituto de Estradas de Portugal (IEP).

- *Introducing bias: data sub-sampling*

I considered two types of sampling effects in the data which lead to environmental bias in the data and which we expect to affect the predictive models: geographical bias and sample size bias. Geographical sampling bias in locational datasets can occur for several reasons such as accessibility, degree of protection or attractiveness of the areas (Romo *et al.* 2006). Indeed, most ecological studies are conducted in protected areas; also, the more accessible or more attractive the areas, the more naturalists and volunteers will be present and collecting data. Sample size bias refers to the shortage of locational data or small datasets, which is frequently the case, due to the cost involved in field data collection. These biased data subsets were generated by sub-sampling the control dataset according to specific criteria, by using probability density functions (PDFs), generated in a GIS. The probabilities assigned in the PDF maps determined the proportion of presence points that were kept from the control dataset, in each of the subsets (Table 2.3).

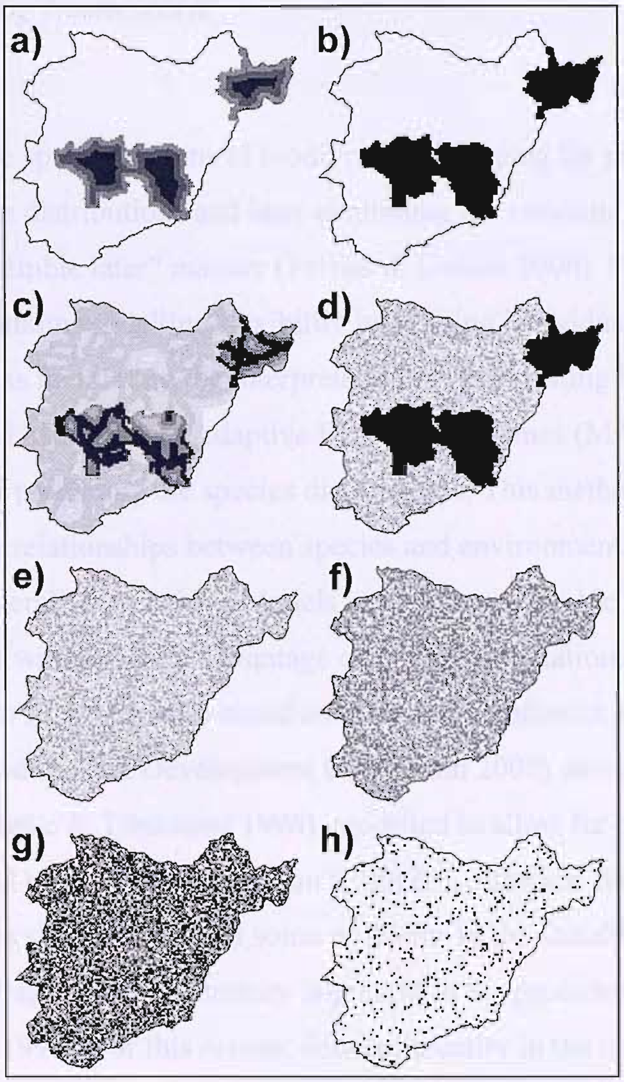
The geographically-biased subsets simulated opportunistic or purposive sampling schemes where the true species prevalence remains unknown. For each class of the PDF layers, I kept an equal number of absences and presences for all species. Taking into account the nature of the study region, with its three SPAs and road network, I generated four different subsets with a decreasing degree of geographical bias (Figure 2.1 a-d).

Table 2.3 - Description of datasets, and respective measures of similarity (average percentage values) with the control dataset: proportion of data sample points retained; and data range overlap in the predictor variables

Dataset	Description	Data retained	Range overlap
No bias			
CONTROL	Complete field-sampled dataset	100	100
Geographical bias			
G1	Maximised sampling in the geographical centres of the SPAs, excluding all data outside	8.74	57.29
G2	Maximised sampling uniformly within the SPAs, excluding all data outside	12.62	64.58
G3	Maximised sampling inside the SPAs, close to roads, and a very low intensity sample of data outside, also maximised close to roads	13.31	64.17
G4	Maximised sampling uniformly within the SPAs and a very low intensity random sample of data outside	16.03	72.54
Sample-size bias			
S1	Very low intensity geographically stratified random sample of the Control dataset	11.66	80.52
S2	Low intensity geographically stratified random sample of the Control dataset	21.49	86.41
S3	Medium intensity geographically stratified random sample of the Control dataset	44.33	91.59

The sample-size biased subsets simulate geographically stratified random sampling schemes (by using randomly distributed PDFs), with a varying degree of sample intensity, this way generating differently sized small datasets (Figure 2.1 e-g). Within these subsets, both presences and absences were equally sampled from the control dataset, according to the probabilities assigned on the PDF layers, so the data prevalence for each species is expected to remain roughly stable at the level it occurs naturally in the area.

Figure 2.1 - Data sampling and sampling bias. Probability Density Function layers generated to produce the respective biased subsets: a) G1 - geographical centres of the SPAs; b) G2 - uniform within SPAs; c) G3 - SPAs and road network; d) G4 - SPAs and very low intensity random; e) S1 - very low intensity random; f) S2 - low intensity random; and g) S3 - medium intensity random. Control data sampling: h) location of the sampling squares (black) in the study area



The similarity between each of the subsets and the control was assessed by two different measures: proportion of data sample points retained; and data range overlap in the predictor variables (Table 2.3). The relationship between these two measures of similarity differed greatly between the geographically biased (G) and random (S) subsets: while a small reduction in sample size in the geographically biased datasets corresponded to a great decrease in the respective proportional range overlap (in the data domain), the opposite was observed on the randomly distributed subsets – a large reduction in sample size corresponded to only a small decrease in data range overlap.

- *Modelling framework*

For modelling the spatial patterns of biodiversity, we opted for predicting individual species distributions and later combining the resulting maps in a "predict first, assemble later" manner (Ferrier & Guisan 2006). By doing so, we allowed for maximum modelling flexibility in defining individual species spatial patterns, as well as facilitating the interpretation of the resulting biodiversity patterns. We used Multivariate Adaptive Regression Splines (MARS) (Friedman 1991) models for predicting the species distributions. This method is capable of fitting non-linear relationships between species and environment, much in the same way as Generalised Additive Models and with comparable modelling performance, but with the great advantage of being computationally less demanding due to its simple rule-based architecture (Leathwick *et al.* 2006). Models were fitted in R (R Development Core Team 2008) using code from the 'mda' library (Hastie & Tibshirani 1996), modified to allow for binary data (logit link function) and model cross-validation (Elith & Leathwick 2007). MARS models have been shown to present some problems in the variable selection procedure when high multi-collinearity is present in the predictor variables (De Veaux & Ungar 1994). For this reason, data collinearity in the training data was inspected, and whenever two variables were correlated with a (Pearson r) value greater than 0.7, one of them was excluded (Freedman *et al.* 1992). The variable retained was the one that scored highest on a Mann-Whitney U rank test between the descriptors and response variables. The models were then applied over the full study area for prediction of the probabilities of occurrence of each species throughout the region, on all datasets (control and subsets).

An ecological interpretation of the models obtained is not presented, as were considered out of the scope of the present paper.

- *Spatial patterns of species occurrence and biodiversity*

Typically, the resulting probabilities from SDMs are converted into binary maps of species distributions, in order to be subsequently assembled into species richness or biodiversity maps (Cumming 2000). However, this requires the definition of a threshold on the output probabilities, this way losing much of the contained information. Moreover, the definition of thresholds of occurrence in species distribution predictions is an issue of large debate (Liu *et al.* 2005). The direct use of the probability of occurrence values, on the other hand, has the advantage of being threshold-independent and this way quantifying the uncertainty of a species occurrence at a location (Loiselle *et al.* 2003). This probabilistic approach has been used successfully in biodiversity mapping, particularly on studies aimed at reserve network design (Polasky *et al.* 2000; Cabeza *et al.* 2004), with significant differences when species occurrence uncertainty was higher (i.e., with probability values furthest from 0 or 1). Hence, we used the predicted probabilities of species occurrence as descriptors of the species distribution patterns and their cumulative probabilities as spatial representations of the steppe bird biodiversity patterns in the region, on all datasets. These biodiversity maps were subsequently re-scaled to values between 0 and 1 to allow a better comparison between different datasets (this way accounting with the possibility of some models not being able to be fitted, particularly in the smallest datasets). This way, a resulting value of 1 would represent an area with probability of occurrence values of 1 on all modelled species.

- *Comparative analysis*

To assess the effects of the different data sampling biases on the distribution models and on the spatial biodiversity patterns, we compared results from the control and biased datasets at both the individual species level and the community (biodiversity) level. At the individual-species level, we inspected the effects both on the habitat models (performance and structure consistency) and on the resulting probability maps. Model performance was assessed using Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) values (Hanley & McNeil 1982), by performing a 10-fold cross-validation, while controlling for prevalence (AUCcv). We assessed model structure consistency between the different data subsets, for each species, by calculating the percentage of variables from the control models retained on each of the respective subset models.

The predicted individual-species probability maps were compared by calculating the mean absolute pixel difference between the respective control and each of the subsets, using the Map Comparison Kit software package (Visser & De Nijs 2006). These comparisons provided pixel-by-pixel measures of similarity / dissimilarity between model predictions, and were therefore used as indicators of consistency in the resulting spatial patterns. Also, this measure is expressed in the same units as the predicted maps, so it reflects the average change in the probability of occurrence values on each map.

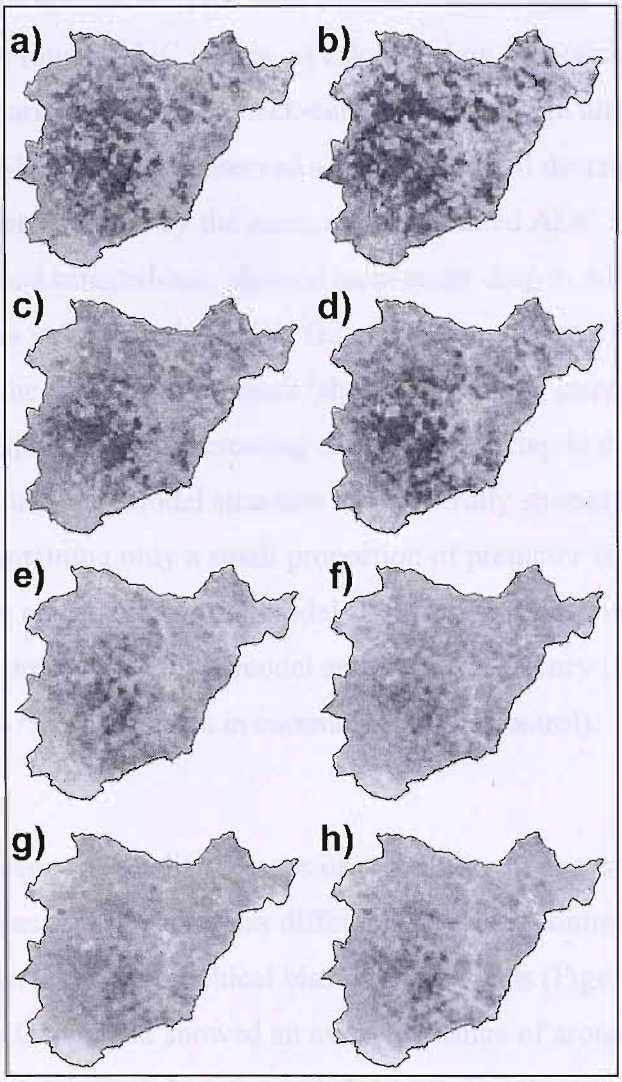
The resulting biodiversity maps were compared using the same dissimilarity measure. In this case, however, the resulting units on this absolute difference (or change) metric is, as in the biodiversity maps, the proportional cumulative probability of occurrence of all species.

2.4. *Results*

- *Control models of species distributions*

The MARS models were fitted for all species in the control datasets with varying degrees of performance (Table 2.1). The AUC values as calculated on the training data ranged from 0.60 to 0.94 (mean 0.79 ± 0.023), and the cross-validated AUC (AUCcv) from 0.52 to 0.78 (mean 0.70 ± 0.021).

Figure 2.2 - a) to h) Biodiversity maps (cumulative probabilities) calculated on the respective datasets: a) G1 - geographical centres of the SPAs; b) G2 - uniform within SPAs; c) G3 - SPAs and road network; d) G4 - SPAs and very low intensity random; e) S1 - very low intensity random; f) S2 - low intensity random; g) S3 - medium intensity random; and h) Control



The predicted probabilities of occurrence for the twelve species were added together in order to compile the control biodiversity map (Figure 2.2 h). On this map, the areas of greater cumulative probabilities which correspond to the core areas for steppe birds in the region (mostly within the Castro Verde SPA) are easily identifiable.

- *Models with geographically biased data sub-samples*

The introduction of geographical bias in the data sampling occasionally had the effect of impeding the successful fitting of the models. This happened on four occasions for three species with the weakest associations in the control models, i.e. those with the lowest AUC values, as calculated on the training data: Crested/Thekla Lark (subset G2); Black-eared Wheatear (G2 and G4); and Stonechat (G3). Moreover, we observed a relatively small decrease in model performance, when assessed by the mean cross-validated AUC values (Table 2.4). Subset G3, the most affected one, showed an average drop in AUCcv of around 10%, whereas this value on subsets G1, G2 and G4 was always smaller than 5%. The variation in the model performance (shown by the SE) increased with increasing geographical bias (decreasing data range overlap in the predictor variables) in the datasets. Model structure was generally strongly affected, the subset models containing only a small proportion of predictor variables in common with the respective control models. Subsets G4, however, being the least biased also presented greater model structure consistency (though with an average of only 47% of variables in common with the control).

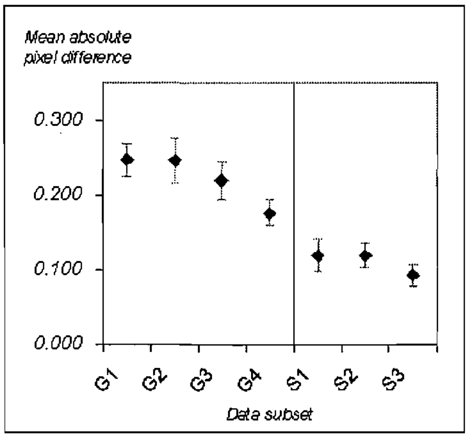
Comparison between the predicted maps of probability of occurrence showed a consistently increasing effect (greater difference between control and subset maps) with an increased geographical bias on the datasets (Figure 2.3). For example, subsets G1 and G2 showed an average change of around 0.25 (SE = 0.022 and 0.029, respectively) on the probability scale, whereas subsets G3, a value of 0.22 (±0.025) and subsets G4 resulted in a mean value of 0.18 (±0.017).

Calculation of biodiversity maps from these four datasets generally resulted in great discrepancies from the control map (Figure 2.2 a to d). In effect, these maps presented great dissimilarity (high mean pixel difference values) from the control, and the core steppe bird areas in the region were no longer identifiable. The observed dissimilarity values were of 0.11, 0.14, 0.10 and 0.10 for datasets G1 to G4, respectively (Figure 2.4). For these models alone, no significant correlation was found between the observed differences on the biodiversity maps and the measures of data similarity (proportion of data retained and data range overlap).

Table 2.4 - Mean and SE values of the model comparisons: cross-validated performance (AUCcv) and structure consistency (percentage variables in common with the control models)

Dataset	Performance	Structure consistency
Control	0.70±0.021	1.00±0.000
G1	0.67±0.038	0.36±0.089
G2	0.69±0.037	0.30±0.075
G3	0.63±0.031	0.38±0.058
G4	0.68±0.025	0.47±0.067
S1	0.57±0.027	0.35±0.081
S2	0.57±0.019	0.49±0.077
S3	0.64±0.026	0.38±0.084

Figure 2.3 - Average and SE values of mean absolute pixel difference between control and subsets, across species, on the predicted individual-species probability maps



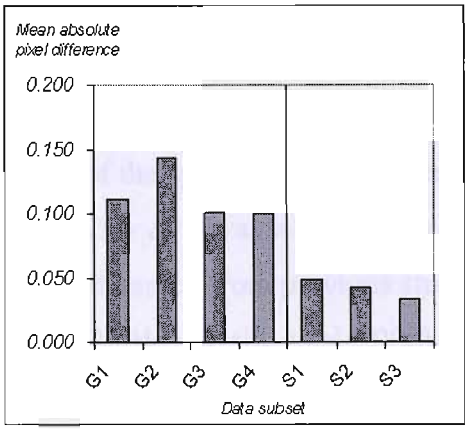
- *Models with sample-size biased sub-samples*

The (geographically stratified) random sub-sampling of the control dataset resulted in considerably poorer performing models than the controls. This observed effect was greater with a decreasing sample effort in the subsets, e.g. a drop in AUC on subsets S1 and S2 of around 18% (however, with smaller SE on the latter), and 9% on subset S3. Model performance was significantly correlated with the proportion of data retained from the control dataset (Spearman rank correlation *rho* or *rs* = 0.35; *n* = 36; *p* < 0.05). As with the geographically biased datasets, model structure was severely affected by data bias, for all subsets (Table 2.4).

Comparison between model predicted maps of probability of occurrence showed a relatively small effect (small absolute pixel difference values) from this type of data sampling bias. Even so, the greater the bias (smaller the dataset), the higher was the dissimilarity on the output maps (Figure 2.3). E.g., subsets S1 and S2 resulted in average changes of 0.12 (SE = 0.022 and 0.017, respectively) and S3 in values of 0.09 (±0.014) on the probability scale. The differences observed in these datasets, however, were always much smaller than those observed from the geographically biased ones.

The resulting community maps were very similar to the control, with relatively low dissimilarity values, even though increasing with the decreasing sample size (Figure 2.4). In fact, the dissimilarity was always less than half of that obtained from the maps generated with the geographically biased datasets (values of 0.049, 0.042 and 0.033, for respective datasets S1 to S3). For these models, the differences observed between the biodiversity maps were strongly, negatively and highly significantly correlated with both data similarity measures, proportion of data retained (*rs* = - 0.87; *n* = 36; *p* < 0.001) and data range overlap (*rs* = - 0.94; *n* = 36; *p* < 0.001). Additionally, identification of the core steppe bird areas in the region was still possible, with this type of bias (Figure 2.2 e to g).

Figure 2.4 - Mean absolute pixel difference values between Control and subsets, on biodiversity maps



- *Data sampling bias effects on model predictions*

Considering all biased subsets together, the observed changes in the individual-species predictions and in the biodiversity maps were always highly significantly correlated with both measures of data bias. Thus, the smaller the datasets used in the models, the larger were the differences from the control on both individual-species (rs = - 0.42; n = 80; p < 0.001) and community predictions (rs = - 0.59; n = 84; p < 0.001). However, these differences in predictions were even more strongly correlated with the average proportion of data range overlap on the predictor variables (rs = - 0.61; n = 80; p < 0.001, on species predictions; and rs = - 0.83; n = 84; p < 0.001, on biodiversity maps).

2.5. Discussion

In this study we investigated the effects of data sample bias on predictive SDMs, particularly when used as a tool for characterising the spatial patterns of biodiversity. The performance of these models and the resulting predictions are ultimately dependent on the locational datasets used. Their size, geographical and environmental coverage, determined by the data sampling, have been reported to affect the modelling process (Elith *et al.* 2002; Stockwell & Peterson 2002;

Kadmon *et al.* 2003; Edwards Jr. *et al.* 2006; Loiselle *et al.* 2008). By introducing geographical bias on a dataset, based on previous knowledge of the study region (location of the regional SPAs and road network), we have simulated a common situation, where due to time, financial or access restrictions (or even a lack of knowledge of the species) data are sampled in a non-random or purposive manner (see e.g. Reddy & Dávalos 2003). It is also often the case that SDMs rely on existing biased datasets, from previous studies or from herbarium and NHC data (Graham *et al.* 2004; Loiselle *et al.* 2008). Depending on the study area, geographical sampling bias is expected to result in a varying degree of environmental bias (variability reduction and skew) of a dataset, which in turn will affect model predictions (Barry & Elith 2006). For example, Kadmon *et al.*(2004) found that geographical sampling bias (close to a road network) in a plant locational dataset still resulted in good model predictions due to the fact that the road network was relatively unbiased in terms of the environmental gradients of the area (i.e., provided a good sample of these gradients). Nevertheless, it is logical that the greater the geographical bias in a dataset, the greater the resulting environmental bias. We demonstrated this in our study by observing the change in data coverage (range overlap in the predictor variables) between the different geographically and by comparison with the non-geographically (random) biased data subsets generated. Even with the use of carefully designed data sampling schemes (geographically or environmentally stratified in a random manner), environmental bias can still occur, particularly when the data sample is too small to describe the full environmental variability present in the study region. The randomly sorted data (sample size biased) subsets of varying sampling intensity used in this study, allowed us to test this effect, by generating datasets with varying sample size and proportional coverage of the region's existing environmental gradients. The degree of environmental bias in these subsets was, nevertheless, generally much smaller than that in a geographically biased dataset of the same size.

Sample size has previously been found to affect the accuracy of SDMs. Stockwell & Peterson (Stockwell & Peterson 2002) found that model accuracy on the training set decreases and on the test set increases with larger samples.

Also Kadmon *et al.* (2003) and Wisz *et al.* (2008), using independent test data to evaluate their models, found this pattern of decreasing model accuracy with decreasing sample sizes. In our study, it was not possible to use an independent evaluation dataset, but a 10-fold model cross-validation was implemented instead. We observed a similar pattern particularly on the randomly sampled subsets – the greater the proportion of data retained (greater the sample size), the greater the model performance. When grouping all cases together (geographically and randomly sampled subsets), however, we achieved inconclusive results regarding the relationship between sample size and model performance. On the other hand, we observed an increase in performance variability (standard error of the mean) with decreasing sample size, across species and between subsets, which agrees with the findings of the two previous studies.

Several authors have also investigated the effects of environmental bias on the performance of distribution models. Kadmon *et al.* (2003) and Edwards Jr. *et al.* (2006), compared random and biased datasets and both concluded that the larger the environmental bias, the lower the model prediction accuracy. In the first of these studies, the authors further concluded that climatic (environmental) bias in the data was more influential than data quantity, and that it could alone explain a large proportion of the variation in predictive accuracy. We found a similar pattern, though only on the geographically biased subsets. On these, the range overlap in the data domain of the predictor variables explained 21% of the observed variability in model performance. We could not, however, find a similar correlation when compiling all cases together. Loiselle *et al.* (2008), on the other hand, found sample size to be more influential on model performance than environmental bias. In their study, though, the authors used presence-only data and a machine learning algorithm (maximum entropy), which might be affected by data bias in a different way than classification tree (Edwards Jr. *et al.* 2006) or regression methods like the one used in our study.

We found that data sampling bias not only affected model performance, but also model structure in terms of the variables selected by the models: different data subsets were fitted on different sets of predictor variables by the MARS variable selection procedure. In fact, on average (across species, on each subset) the subset models had fewer than half of the variables in common with the control models. Moreover, this pattern was observed even when using a randomly sorted dataset that covered a large proportion of the control dataset, as in the case of subset S3 (where the data range overlap is 91.59%). A similar conclusion was reached by Edwards Jr. *et al.* (2006); the overlap in the predictor variables selected by their models, between biased and non-biased datasets was always below 38%. Recent research by Osborne & Leitão (*in press*) (see Appendix A.2) has also found large inconsistencies in model structure when small amounts of locational error are introduced in the data. This suggests that the variable selection algorithms of SDMs are highly sensitive to any source of error, and ecological inference about species habitat selection should be made with extreme care. We therefore recommend that these interpretations should only be taken when using balanced (unbiased) and large datasets. On these cases, nevertheless, a reliable ecological interpretation from such models could still be attained by, e.g. running data randomisations together with an evaluation about the model structure stability.

Analysis of both the predicted probability maps and resulting biodiversity maps showed a clear pattern of an effect of data sampling bias on the resulting predictions. Indeed, both measures of data similarity with the control dataset used in this study were highly correlated with the differences observed in the predicted patterns of species occurrence and biodiversity. Environmental bias, however, was more influential on the prediction patterns than sample size bias, and the mean data range overlap (in the predictor variables) alone could explain a great proportion of the observed differences in species probability and biodiversity maps (according to the Spearman rank correlation rho scores). A recent study by Grand *et al.* (2007) focusing on reserve selection algorithms, typically based on optimisation of biodiversity measures such as species richness or complementarity, found a similar pattern. These authors also compared the results of a presumably complete locational dataset (of Proteaceae in South Africa) with those of geographically (road) biased and random subsamples in order to assess the impacts of sampling bias and sampling effort on reserve selection results. Their conclusions were that data sampling bias impacted the predictions more severely than decreased sampling effort alone, concurring with our study.

- *Concluding remarks*

SDMs are a valuable tool for many different ecological applications as they provide information on species and biodiversity spatial patterns which would otherwise be mostly unavailable. Locational datasets on which they rely are, however, frequently unbalanced as a result of data sampling bias. This, in turn, has profound impacts over all phases of the modelling process: performance, interpretation and predictions. Data sampling schemes which generate greater environmental bias (less representative coverage of the environmental gradients), such as geographically-biased or purposive sampling will have a greater effect on the distribution models. Special attention should be paid to the quality of existing datasets as well as the sampling design for collection of new data. Ideally, statistical tests should be performed on the data, to detect the presence of bias before modelling is undertaken, and any interpretation of models based on biased data should be made with explicit awareness of the existing bias.

2.6. Acknowledgements

This work benefited from a research grant by FCT (SFRH/BD/12569/2003) and support from Liga para a Protecção da Natureza (LPN), Sociedade Portuguesa para o Estudo das Aves (SPEA), Instituto de Estradas de Portugal (IEP) and Perímetro Florestal da Contenda. Supplementary data was provided by Pedro Rocha, Ana Delgado and Inês Henriques and projects EDIA/PMo5.4, PRAXIS/C/AGR/11062/1998, PRAXIS/C/AGR/11063/1998, and LIFE02/NAT/P/8476. The R scripts used for fitting the MARS models and crossvalidations were kindly provided by Jane Elith. Maria João Santos provided helpful comments on an earlier version of this manuscript.

3. *Analysis of remote sensing data to derive predictor variables for use in species distribution models*

3.1. *Introduction*

A range of advanced statistical tools for predicting species distributions are increasingly becoming available to ecologists and environmental managers (Guisan & Zimmermann 2000; Guisan & Thuiller 2005). Their methodological principle lies within ecological niche theory, in which the distribution of a species is determined by the extent of its niche and is therefore inferred by modelling the respective habitat distributions (Guisan & Zimmermann 2000; Austin 2007). These methods couple incomplete species locational data (response variable) with environmental descriptors (predictor variables), and are capable of predicting species distribution patterns over large areas, as well as providing insights into the underlying species-environment associations.

Remote sensing data, by thoroughly and systematically describing the Earth's surface, are an excellent source of detailed environmental data to be used as predictors in species distribution models (SDMs) (Kerr & Ostrovsky 2003). Moreover, they enable the collection of data in remote and otherwise inaccessible areas. It is not the aim of this work to present a review on the use of RS data for biodiversity mapping, as there are some extensive reviews already present in the scientific literature (Nagendra 2001; Gottschalk *et al.* 2005; McDermid *et al.* 2005; Leyequien *et al.* 2007; Gillespie *et al.* 2008). We intend, however, to describe some common sources of RS data, their characteristics and usage, as well as to compare available data analysis approaches for feature extraction in order to subsequently use them in SDMs. These approaches are compared in terms of achieved model performance, and respective model generality and interpretability.

- *Predictor variables*

Environmental predictors can be described in terms of their position in the chain of processes that link them to their impact on the species. Thus they may be either proximal or distal, proximal variables being those that are causal, determining the species response, while distal variables are those that only indirectly affect the species (Austin 2002). From another perspective, one can identify three main types of environmental variables, based on their known biophysical processes: resource, direct and indirect variables. Resource variables refer to matter and energy consumed by the organisms, like water, food, etc.; direct variables are environmental parameters with physiological importance, but that are not consumed, like temperature, humidity or pH; and indirect variables are those that have no direct physiological relevance for a species, like elevation or habitat type (Austin 1980; Guisan & Zimmermann 2000; Austin 2002).

The arrangement of predictors into direct/indirect or proximal/distal directly relates to model generality. A model which uses only indirect or distal variables lacks generality and only has local value for both prediction and understanding, and hence should not be applied to large areas nor to elsewhere (Austin 2002). This is because these variables reflect combinations of different resources and direct gradients specific to the study location. On the other hand, they are easily measured in the field and usually show good correlations with the observed species patterns (Guisan & Zimmermann 2000; Austin 2007). Another possible advantage of using indirect variables such as land use or habitat type is that they can directly relate to land management practices and thus aid local conservation efforts. Models based on proximal resource and direct gradients are the most robust and widely applicable (with more generality), but because these variables are the most difficult to measure they are often impractical for use (Austin 2002).

Hence, the choice of the predictor variables to use must be primarily dependent on the aim of the study, whether it is just intended to predict the species distribution within the study region and for the period of study, if the intention is to predict over different conditions (other area or other period in time), or if ecological interpretations of the model results are to be drawn. Other determining factors relate to the existing knowledge of the species ecology, and the ability to measure relevant environmental descriptors. RS data offer a suite of different data sources and processing techniques, capable of measuring the existing environmental gradients, which can then be used as predictors in these models.

- *Remote sensing data*

Earth observation satellites deliver data with spatial resolutions ranging from under 1 m to up to 8 km (Table 3.1). Passive remote sensors collect data on surface radiation, most commonly reflected sunlight (optical sensors), but also emitted radiation (thermal). Optical sensors measure the reflected radiation either in a single panchromatic spectral band covering all the visible light wavelengths (0.4-0.7 μm), or in distinct spectral bands (multi-spectral), each covering specific portions of the electromagnetic spectrum. Typical spectral regions include the visible light, near-infra-red (0.7-1.4 μm), short-wave infrared (1.4-3 μm) or mid-wave infrared (3-8 μm), each used to adequately describe specific surface features. For example, terrestrial vegetation can be well described by band combinations on the red and near infrared wavelengths, many times used in the form of vegetation indices (Tucker 1979), whereas soil moisture is well described by combinations of bands on the near and short-wave infrared portions of the spectrum (Wilson & Sader 2002; Haubrock *et al.* 2008). Thermal sensors measure the emitted radiation on the long-wave or thermal infrared (8-15 μm) region of the spectrum and are capable of describing surface temperature patterns. Active remote sensors, on the other hand, emit energy to the surface (at specific wavelengths) and then record the backscattered signal, this way calculating some target's characteristics such as location, height or orientation (Dobson *et al.* 1995). These are suited to produce detailed topographic maps, but can also retrieve information on surface moisture and roughness, forest canopy structure or agricultural crop structure, yield and orientation.

Table 3.1 - Commonly used satellite remote sensors, ordered by pixel spatial resolution. (Legend: CAR - Cartography; LCM - Land cover mapping; MET - Meteorology & Climatology; NAT - Natural resources monitoring; OCE - Ocean monitoring; VEG - Vegetation & agricultural crop monitoring; TOP - Topography)

Sensor	Type of sensor	Pixel spatial resolution (m)	Applications
GOES	Optical and thermal	1000, 4000, 8000	MET
Meteosat MVIRI/SEVIRI	Optical and thermal	1000, 2500, 5000	MET
SPOT VGT	Optical	1150	NAT; VEG
NOAA AVHRR	Optical and thermal	1090	MET; OCE; NAT; VEG
ERS ATSR	Optical and thermal	1000	OCE; NAT; VEG
Terra MODIS	Optical and thermal	250, 500, 1000	MET; OCE; LCM; NAT
ENVISAT MERIS	Optical	300	MET; OCE; LCM; NAT; VEG
ENVISAT ASAR	Radar	30, 150, 300	TOP; OCE; NAT
Landsat TM/ETM+	Optical and thermal	15, 30, 60, 120	LCM; NAT; VEG
RADARSAT-1 & -2	Radar	3, 8, 10, 12, 25, 26, 30, 50, 100	NAT; VEG
NASA SRTM	Radar	30, 90	TOP
Terra ASTER	Optical and thermal	15, 30, 90	LCM; NAT; VEG; TOP
IRS LISS-III & -IV	Optical	5.8, 23.5, 70	LCM; NAT; VEG; CAR
CHRIS-PROBA	Optical	20, 34	LCM; NAT; VEG
ERS SAR	Radar	30	TOP; NAT; VEG
EO-1 Hyperion	Optical	30	LCM; NAT; VEG
SPOT HRV/HRVIR/HRG	Optical	2.5, 10, 20	LCM; NAT; CAR; TOP
TerraSAR-X	Radar	1, 3, 16	NAT; VEG
RapidEye	Optical	6.5	NAT; VEG; CAR
Ikonos	Optical	0.82, 3.28	LCM; NAT; VEG; CAR
QuickBird	Optical	0.61, 2.44	LCM; NAT; VEG; CAR
GeoEye-1	Optical	0.41, 1.64	LCM; NAT; VEG; CAR

The applications of these data are multi-fold, depending on their characteristics (the type of information collected and its spatial resolution), which range from global scale climatologic and meteorological studies; to natural resources, vegetation and agricultural crop monitoring at regional and landscape scales; and to detailed cartographic and topographic characterisation. Also, recent advances in technological research has permitted a multiplication of these data sources including the emergence of very-high resolution imagery satellites, such as QuickBird or GeoEye-1 (which deliver surface reflectance measurements at sub-metre resolutions); the development of sensors with dedicated spectral bands for the correction of atmospheric effects (reducing an important source of error in the characterisation of the Earth's surface), like the ENVISAT MERIS; and the development of spaceborne hyperspectral sensors (capable of collecting information on hundreds of spectral bands from across the electromagnetic spectrum), like the EO-1 Hyperion or the prospective EnMAP satellite (due to be launched in 2012).

Given this, the selection of the RS data sources for use in SDMs must be carefully made, and the decision should be guided by the aims of the study, the sensor characteristics and its capability of describing environmental factors that reflect the species occurrence patterns at the desired scale of study.

- *Digital image processing methods for deriving predictor variables*

The multitude of digital image processing (DIP) methods to analyse RS data can hinder the decision on how to best derive useful predictor variables for use in species distribution modelling. Some products are available in a ready-processed format which can be directly used as predictors in SDMs, such as the SRTM-derived Digital Elevation Model, or the Global Land Cover map provided by the Global Land Cover Facility (Hansen *et al.* 2000), but this is often not the case. Also, by carefully selecting and analysing a specific RS image, it is possible to generate fit-for-purpose predictor variables, which the (ready-processed) standard products are not able to provide. It is therefore important to assess the available methods for processing these data.

Perhaps the most commonly used approach is to classify the imagery into land use/land cover or habitat classes, and then derive measures of composition (proportion of cover for each class) and configuration (like fragmentation or connectivity) for use as predictors in distribution models (Austin *et al.* 1996; Tucker *et al.* 1997; Luoto *et al.* 2002). Image classification can be achieved by means of unsupervised or supervised methods. Unsupervised classification algorithms are able to aggregate the imagery into "natural" clusters of separable spectra (spectral classes), which can then be associated to specific land-cover classes or habitats. Typically, iterative methods are used, like the ISODATA - Iterative Self-Organizing Data Analysis Technique (Ball & Hall 1965) or the SOM - Self-Organizing Map (Kohonen 1990). Supervised methods, on the other hand, require the prior identification of areas with known classes, which are used to train the classifier to then be applied to the remainder of the image. The latter methods permit the classification of the imagery into desired classes, i.e. those considered important to describe the species occurrence patterns. However, the final classification is highly sensitive to the identification of the training areas, which is prone to human error. The most popular of the supervised classification methods is the Maximum Likelihood Classifier (MLC), a parametric classifier which makes use of the classes' mean spectral values and respective covariance matrices to assign a probability of membership of the unknown pixels to each class, and subsequently label it with the most likely category, i.e. to the class with the highest probability. Further developments in DIP have been the introduction of (non-parametric) machine learning methods, such as Artificial Neural Networks (ANN) (Benediktsson *et al.* 1990) or Support Vector Machines (SVM) (Huang *et al.* 2002a) for image classification in order to overcome the problems of data distribution constraints, allowing the usage of data with complex (multimodal) distributions as well as nominal or ordinal data.

These methods produce a hard classification of imagery (i.e. the assignment of each pixel to a specific class), which generates discrete classified maps. While this is a common procedure, it does not accurately describe the vegetation continua of gradually integrating classes, nor the case of mixed pixels, i.e. pixels containing more than one class (Wood & Foody 1989; Smith *et al.* 1990). Soft classification techniques, such as probability mapping (as derived from e.g., the MLC or the Mahalanobis distance), fuzzy-c-means or non-parametric methods, such as the ANN, have been used to this end (Foody 1996). Another related method is the spectral unmixing of imagery pixels into its cover fractions (Lennington *et al.* 1984; Adams *et al.* 1986). This can typically be achieved through the use of a Linear Spectral Mixture Model (LSMM), also called linear Spectral Mixture Analysis (SMA), which assumes that the spectral signature of a given pixel is the result of the linear mixture of the spectra of its component features (Settle & Drake 1993). It does, however, require the definition of distinct (and linearly independent) spectral end members that might not match the desired thematic classes, as well as ignoring the non-linear mixture effects resulting from multiple scattering of radiation among different target materials. Non-linear spectral un-mixing has been suggested for overcoming these problems (Foody *et al.* 1997). These methods would also permit the definition of habitat classes with complex spectra, and therefore suit the intended application, but up till now they remain largely untested.

Soft image and sub-pixel classification techniques are not commonly used methods for deriving SDM predictors. On the other hand, spectral indices (band combinations and ratios) such as those of vegetation and moisture/wetness have wide popularity in SDM research (Wallin *et al.* 1992; Osborne *et al.* 2001; Zimmermann *et al.* 2007). Time-series of such indices are capable of describing phenological processes (Reed *et al.* 1994; Jakubauskas *et al.* 2001) and can be particularly useful for describing species distributions (Osborne *et al.* 2001; Leitão *et al.* 2006). Additionally, a few studies have made use of texture measures for assessing habitat heterogeneity to be used as a predictor of species distributions (Hepinstall & Sader 1997; Imhoff *et al.* 1997). Also, the raw image reflectance values, by describing surface characteristics which can relate to specific environmental conditions, can be used directly as predictor variables (Hepinstall & Sader 1997; Zimmermann *et al.* 2007). The multitude of DIP techniques available further highlights the vast potential of RS data for deriving useful SDM predictor variables.

- *Case study*

A case study is presented, where the distributions of 13 steppe bird species in an agricultural landscape in southern Portugal are modelled. In this study different DIP techniques for deriving model predictors are compared, and relevant issues related to RS data analysis are further discussed. For ease of interpretation we use a single source of imagery, and compare single-method approaches, even though a combination of different data sources and different methods for information extraction is more likely to generate good distribution models (Zimmermann *et al.* 2007; Buermann *et al.* 2008). Hence, data from the Landsat Thematic Mapper (TM) sensor was used, which is the most commonly used source of imagery in ecological applications (Cohen & Goward 2004). The chosen data relate well to the specific research question as they are capable to describe relevant habitat classes at an adequate scale. The observed landscape dynamics are described, in order to help understand the important factors that can influence the distribution of the species in the area. The choice of the data analysis methods are discussed in terms of their advantages and limitations, as well as the type of predictor variables generated and the resulting model characteristics in terms of performance, interpretability and generality.

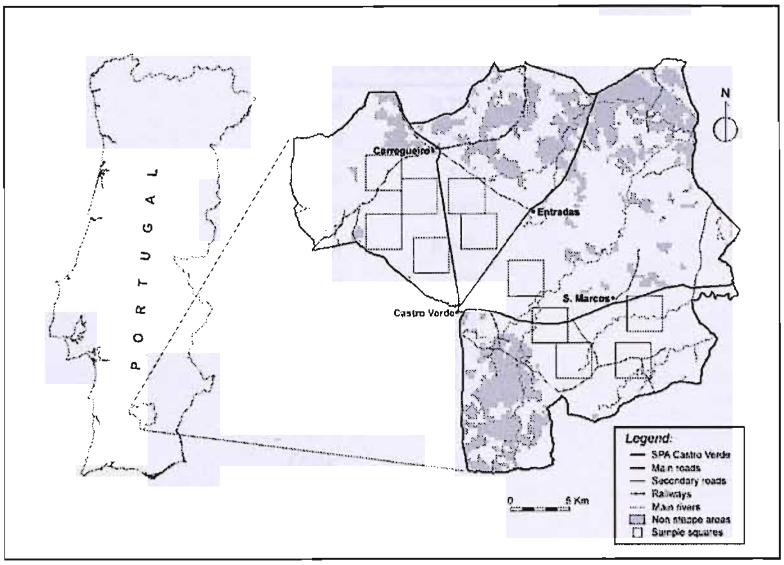
3.2. Methods

- *Study area*

The study was conducted within 11 randomly selected squares of 3x3 km, fully within pseudo-steppe habitats inside the Castro Verde SPA (Figure 3.1). (see Chapter 1.2 for a general description of the Castro Verde SPA)

The cereal pseudo-steppes are a spatio-temporal mosaic resulting from low-intensity agricultural practices (Moreira *et al.* 2007) (see Appendix A.1). Traditionally, and as a result of low productive soils, a rotational system is used, where each farm is divided into parcels. Each parcel is cultivated with winter cereal crops (mostly wheat *Triticum* spp. and oats *Avena* spp., sometimes mixed forage crops) for two consecutive years, which become stubbles after harvest, after which the land is left fallow, normally for a period of 2 to 3 years (but sometimes for periods of up to 7 years). The land is then ploughed (before seeding) and the rotation cycle re-initiated. This generates a landscape mosaic dominated by fallow fields (usually 50% of the area or more) which are usually used as pasture for sheep and, more rarely, cattle (Moreira 1999; Delgado & Moreira 2002). These fallows have a diverse floristic composition, including grasses (Gramineae), composite flowers (Compositae), cloves and other legume plants (Leguminosae). Also present are areas of shrublands, sometimes interspersed with old fallows as a result of land abandonment and scrub encroachment. Other less common land uses include afforestations of eucalyptus, umbrella pines and holm oak (Moreira *et al.* 2005), even though these areas were mostly excluded from the sample squares.

Figure 3.1 - The study area, with the main cartographic features, the eleven sample squares and (in grey) the non steppe areas, adapted from the CORINE Land Cover 2000 map



- *Species data collection*

The selection of the sample squares was based on the Corine Land Cover 2000 map (of the EEA), at a spatial resolution of 250 m (land cover classes 2.1.1, 2.3.1 and 2.4.3). For each of the sample squares, a systematic grid of 10 x 10 points was imposed, totalling 1100 sampling points. During the spring of 2006 (between the 20th of March and the 12th of May, covering the birds breeding season), bird censuses were carried out at these sampling points using point counts (circular-plot censuses) with a 5-minute duration and 125 m distance limit and all (visual and auditory) bird observations within the buffer were registered (Fuller & Langslow 1984; Bibby *et al.* 2000). All bird counts were carried out during the birds'period of peak-activity, i.e. the early mornings (first four hours after sunrise) and evenings (last two hours before sunset). From these counts, the occurrence (presence/absence) status of 13 species (Table 3.2) was determined. The Crested Lark *Galerida cristata* and Thekla Lark *Galerida theklae* were categorised to the genus level due to difficulties in reliably identifying all individuals of these two species in the field (Moreira *et al.* 2007) (see Appendix A.1). These data were integrated in a Geographic Information System (GIS), together with the RS data.

Table 3.2 - List of species studied and respective frequency of occurrence in the area

Species	Freq. of occurrence
Milcal	0.780
Melcal	0.283
Galsp	0.230
Tettet	0.162
Cisjun	0.115
Saxtor	0.108
Calbra	0.101
Aleruf	0.077
Cirpyg	0.064
Antcam	0.054
Otitar	0.051
Oenhis	0.033
Buroed	0.031

- *Remote sensing data description and image pre-processing*

We acquired two overlapping Landsat TM full scenes (path/row: 203/34), which cover the whole of the study area. The image selection accounted for the absence of cloud contamination over the area, and the temporal coincidence with the field surveys. Thus, they are dated from 6th of March and 9th of May 2006, this way allowing for the description of the vegetation phenology (and crop cycle) during the season, particularly relevant in agricultural landscapes (Reed *et al.* 1994; Hill & Donald 2003). Data acquired in the six reflective TM spectral bands were used, from both images, which were subsequently pre-processed. The data pre-processing included a radiometric correction into reflectance values (by using post launch sensor coefficients), an image-based Dark Object Subtraction (DOS) atmospheric correction (Chavez Jr. 1988), and a geometrical rectification. The latter was done through the careful identification of ca. 100 ground control points (GCPs) per image from topographic maps at a spatial scale of 1:25000 and applying a bilinear interpolation resampling, maintaining the 30 m pixel resolution, and with a resulting Root Mean Square error of ca. 13.5 m, on both images. Owing to the very low terrain variability present in the area, no topographic normalisation was performed to the imagery.

- *Digital Image Processing*

For an assessment of the advantages and limitations of different DIP methods for deriving good SDM predictor variables, we applied four distinct DIP techniques: hard supervised classification (*Hard*); soft supervised classification (*Soft*); calculation of spectral indices (*SIs*); and spectral band ordination (*Ord*). Additionally, we used the raw spectral bands without any further processing (*Raw*). The first three methods correspond to knowledge-based approaches as they aim to produce ecological meaningful predictor variables. Approaches *Ord* and *Raw* are numerical-based, which keep the numerical complexity (variability) of the data but carry no clear ecological interpretation.

The hard supervised classification (*Hard*) was done with SVM models, generally considered to be a superior method for multiclass image classification (Huang *et al.* 2002a; Foody & Mathur 2004). SVM is a machine learning algorithm based on statistical learning theory, effectively an optimal margin classifier which seeks to find the optimal separating hyperplane between different classes (Boser *et al.* 1992; Cortes & Vapnik 1995). By being non-parametric, and thus not assuming any specific data distribution, is well suited for habitat classes with complex and otherwise hardly separable spectral signatures. We used an improved version of the 'imageSVM' tool (Janz *et al.* 2007) which makes use of the freely available 'LIBSVM' library (Chang & Lin 2001). We used Gaussian radial basis kernel functions to solve the non-linear problems. Model training requires the definition of two parameters, γ that controls the width of the kernel, and the regularization parameter C (C-SVM formulation) which controls the trade-off between maximizing the margin and penalizing the training errors. The multi-class posterior probabilities estimation was based on a one-against-one approach (Hsu & Lin 2002) through the calculation of the normalized sum of the respective pairwise posterior probabilities derived from the SVM decision values (Platt 2000; Lin *et al.* 2007). The image classification procedure then assigned each pixel to the most likely class of membership (with the highest posterior probability). The search for the best C and γ parameter combination was optimised through a 10-fold cross-validation procedure (Belousov *et al.* 2002; Steinwart 2003) and according to the highest average Cohen's Kappa statistic (Cohen 1960) calculated on the respective confusion matrices. This statistic was chosen in order to account for classes with different sizes in the training data. On both phases of model selection and training we used a low termination criterion tolerance ($\varepsilon = 0.0001$), in order to allow for high numerical precision. In order to ensure model numerical stability, the model training data (spectral bands) were previously rescaled to values between 0 and 1. Before this procedure, however, they were all stacked together, this way keeping the relative magnitude between reflectance values of different bands and of different dates.

Table 3.3 - Habitat classes defined for the hard classification of the imagery

Class	Description
March habitat-related classes	
Bare soil	All areas with bare soil or no vegetation, including ploughed fields, but also dirt tracks and paved/built-up areas
Low vegetation	All areas with low vegetation, including both pastured fallow fields and (late or slow growing) cereal / forage crops
Fallow	All fallow fields, excluding those included in "Low vegetation"
Cereal	All cereal and forage crop fields, excluding those included in "Low vegetation", possibly including some grass-dominated fallows
Woodland/shrubs	All areas of woodland and shrubs
Water	All water bodies, including rivers and lakes / dams
May habitat-related classes	
Bare soil	Same as for March
Low vegetation	Same as for March
Fallow	Same as for March, except areas included in "Green vegetation" and "Dry vegetation"
Cereal	Same as for March, except areas included in "Green vegetation" and "Dry vegetation"
Dry vegetation	Areas with particularly dry (senescent) vegetation, mostly herbaceous vegetation (Gramineae) such as cereal / forage fields or grass-dominated fallows
Green vegetation	Areas with particularly green vegetation, mostly in well irrigated areas and along water lines and topographic depressions
Woodland / Shrubs	Same as for March
Water	Same as for March

The classification training data were extracted by careful interpretation of the field notes taken for the 1100 sample points visited during the bird census, and by visually interpreting the imagery. These locations were matched with the image pixels in the GIS, and the specific habitat-related class was determined for only those pixels that fell within a 50 m radius from the visited locations, this way reducing possible labelling errors. When a pixel was composed of more than one habitat-related class it was not included in the training data, thus only pure pixels were used (2926 in total). The classes were defined according to their potential relevance for describing the species occurrence patterns (Table 3.3). The classes *Fallow* and *Cereal* correspond to the two main land cover classes that constitute the steppe mosaic. It was expected, however, that some grass-dominated (Graminea) fallow fields could be classified as *Cereal*, due to the similar reflectance pattern of these vegetation types. The class *Bare soil* included all non-vegetated areas: even though this could also include paved or built-up areas, in the context of the sample squares (within the steppe mosaic), it mainly corresponds to ploughed fields and possibly some bare areas around the farm houses. The presence or abundance of these land cover classes has been previously associated with the studied species, in the region (Delgado & Moreira 2000; Moreira *et al.* 2007) (see Appendix A.1). Vegetation vertical structure, such as height, is also an important descriptor (Moreira 1999), so low vegetated pixels (regardless of the land use) were incorporated in the *Low vegetation* class. Woodland and shrubs, although with a somewhat different ecological interpretation (Moreira *et al.* 2007) (see Appendix A.1), have a similar spectral response (at the spectral resolution of Landsat) thus were considered in one single class *Woodland / Shrubs*. The class *Water* described all the water bodies present in the area. Along the period of study (March to May), the landscape was subject to sharp phenological events, such as the advanced senescence of (mostly) herbaceous vegetation (fallow grasses and forage/cereal crops) in the dryer areas, or its contrasting "greenness" in the more irrigated areas (in topographic depressions, water lines, etc.). For this reason, we considered two phenological classes in May: *Green vegetation* and *Dry vegetation*. The rarity of the classes *Woodland / Shrubs* and *Water* in the study area (within the pseudo-steppe mosaic), however, presented an image classification problem. The low number of pixels from these classes within the training data (34 *Woodland / Shrubs* pixels and 12 *Water*) resulted in a poor spectral characterisation of these classes, and therefore we included 45 additional training pixels (30 and 15 for the respective classes) from known locations outside the sample squares. The classification of the imagery into separate March and May classes, besides fully characterising the landscape on both dates, also allowed us to possibly infer about the timing of important events in the breeding season of individual species. Nevertheless, we found that the "knowledge" of the spectral information from both images greatly improved each of the two individual classifications (see results section), as different habitat classes have different phenological patterns along the season. This way, both classifications, used the same multi-date spectral data (12 spectral bands: 6 from March and 6 from May), differing only on the training data labels. This was also expected to improve the comparison between the two classified maps and its interpretation in terms of temporal (and phenological) change, by reducing errors resulting from the use of different datasets. These errors could include slight image misregistration or the mismatch in the calculation (by the SVM classifier) of the hyperplanes between similar class pairs (for example bare soil and water) due to data scaling issues rather than to real changes in the class characteristics. Classification performance was assessed through inspection of the overall classification accuracy (OCA), the classwise producer and user accuracies and the Kappa statistic (Congalton 1991; Foody 2002), as calculated on a 10-fold cross validation confusion matrix. The additional training data (*Woodland / Shrubs* and *Water* classes) were not considered for the performance assessment because, due to their purposive/non-random sampling, they would not allow an adequate interpretation of the performance statistics (Foody 2002). As a final step, in order to characterize the presence/absence of each habitat class, the classified images were converted into single-class Boolean images.

The soft supervised classification (*Soft*) was done using the same methodological approach as above. However, in this case we used the SVM decision values as measures of class probabilities (or probabilities of membership of pixels to the classes). We considered them to be more suitable than the posterior class probabilities (as previously calculated in the image hard classification) because the latter, by being based on sigmoid functions, can encourage the hard allocation of classes (Foody 1996). The possible advantage of the soft classification approach is that it should allow for an ecological interpretation of the species habitat preferences (in terms of land use practices) while keeping the numerical heterogeneity of the spectral data. Soft classification outputs can be related to the respective pixel fractions of cover, in a spectral unmixing manner (applicable to classes which describe distinct land uses, such as bare soil, fallow or cereal). However, for this purpose the defined classes must be separable in the unmixing space and thus we excluded the class *Low vegetation*. This class is by itself composed of a mix of spectra between the dominant bare soil cover and its low density vegetation, so it would fall within the spectral mixture space of the class *Bare soil* and the class of the respective vegetation (*Fallow* or *Cereal*). Otherwise, all classes were defined as for *Hard* (Table 3.4).

Table 3.4 - Habitat classes defined for the soft classification of the imagery

Class	**Description**
March habitat-related classes	
Bare soil	All areas with bare soil or nor vegetation, including ploughed fields, but also dirt tracks and paved/built-up areas
Fallow	All fallow fields
Cereal	All cereal and forage crop fields, possibly including some grass-dominated fallows
Woodland/shrubs	All areas of woodland and shrubs
Water	All water bodies, including rivers and lakes / dams
May habitat-related classes	
Bare soil	Same as for March
Fallow	Same as for March, except areas included in "Green vegetation" and "Dry vegetation"
Cereal	Same as for March, except areas included in "Green vegetation" and "Dry vegetation"
Dry vegetation	Areas with particularly dry (senescent) vegetation, mostly herbaceous vegetation (Gramineae) such as cereal / forage fields or grass-dominated fallows
Green vegetation	Areas with particularly green vegetation, mostly in well irrigated areas and along water lines and topographic depressions
Woodland/shrubs	Same as for March
Water	Same as for March

On the other hand, the probability of membership of a pixel to the phenological classes (*Green vegetation* and *Dry vegetation*) should instead relate to the phenological stage of the respective pixel's vegetation, rather than to its fractions of cover. In the classification of the March and May habitat-related classes we used 2374 and 2653 training pixels, respectively (the difference between the sizes of these datasets relates to the amount of "mixed" low vegetated pixels excluded, at any particular date). The classification performance was assessed the same way as with the *Hard*. Even though these performance measures are better suited for hard classification approaches (Gómez *et al.* 2008), they should still be good indicators of the SVM model performance. The SVM decision value images for each class were extracted, to serve as model predictor variables.

The third approach used was the calculation of spectral indices (*SIs*), separately for March and May images. By browsing in the RS literature it is possible to find many different spectral indices (band combinations, ratios and normalized differences, etc.) which are commonly used to describe vegetation vigour, soil moisture content or soil brightness. In the search for the best set of indices that could describe the distribution patterns of the studied species, we calculated several candidate (potentially useful) indices (Table 3.5). The Tasselled Cap transformation (Kauth & Thomas 1976; Crist & Cicone 1984) is a band ordination technique designed for reducing Landsat spectral information into three meaningful and (linearly) non-correlated components. These are associated with soil brightness, vegetation greenness and soil moisture, and were originally conceived to describe agricultural crop development. We calculated the Tasselled Cap indices using the standard default coefficients for use with Landsat TM data (Crist & Cicone 1984). In order to describe bare soil, we also calculated the Normalized Difference Soil Index (NDSI) as defined by Rogers & Kearney (Rogers & Kearney 2004). Green vegetation was further characterised by two more indices: the Normalized Difference Vegetation Index (NDVI) (Rouse *et al.* 1973), possibly the single most commonly used spectral index; and the Optimized Soil-Adjusted Vegetation Index (OSAVI) (Rondeaux *et al.* 1996), referred as a better method for describing vegetation in an agricultural context by reducing the effect of the soil background. Also, two additional indices relating to soil moisture content were calculated, the Normalized Difference Moisture Index (NDMI) (Wilson & Sader 2002) and the Normalized Soil Moisture Index (NSMI) (Haubrock *et al.* 2008). Finally, we calculated one index associated with senescent vegetation, the Normalized Difference Senescent Vegetation Index (NDSVI) (Qi & Wallace 2002), which could be particularly useful for describing the vegetation phenological condition in May. There is, however, some overlap in the architecture (for example, the NDSI is the inverse of the NDMI) and rationale of some of the indices, which should result in high collinearity between them as well as similar ecological interpretations (as for example, several indices describing green vegetation). Nevertheless, this issue was addressed at a later stage and is described in the next section.

Table 3.5 - Spectral indices calculated to describe bare soil, green and dry (senescent) vegetation and soil moisture, in both March and May images (Legend: TM3 to TM5 and TM7 – Thematic Mapper spectral bands 3 to 5 and 7)

Description	Formula
Soil Indices	
TC Bright - Tasselled Cap Brightness Index	See Crist & Cicone (1984)
NDSI - Normalized Difference Soil Index	(TM5-TM4)/(TM5+TM4)
Green vegetation	
TC Green - Tasselled Cap Greenness Index	See Crist & Cicone (1984)
NDVI - Normalized Difference Vegetation Index	(TM4-TM3)/(TM4+TM3)
OSAVI - Optimized Soil-Adjusted Vegetation Index	(TM4-TM3)/(TM4+TM3+0.16)
Dry vegetation	
NDSVI - Normalized Difference Senescent Vegetation Index	(TM5-TM3)/(TM5+TM3)
Moisture	
TC Wet - Tasselled Cap Wetness Index	See Crist & Cicone (1984)
NDMI - Normalized Difference Moisture Index	(TM4-TM5)/(TM4+TM5)
NSMI - Normalized Soil Moisture Index	(TM5-TM7)/(TM5+TM7)

Spectral band ordination (*Ord*) was done through a Principal Components Analysis (PCA) (Pearson 1901). This procedure projects the data into an orthogonal space, to derive linearly non-correlated principal components, being particularly useful to handle highly collinear spectral data (Jenson & Watz 1979). Thus, it has been extensively applied to RS data for several ends, such as feature definition and discrimination (Santisteban & Muñoz 1978; Ceballos & Bottino 1997), image classification (Patterson & Yool 1998) or the assessment of multi-temporal and agricultural/phenological patterns (Panigrahy & Sharma 1997; Coppin *et al.* 2004; Lasaponara 2006). The procedure calculates the principal components based either on the data's variance/covariance matrix (unstandardized PCA), or on the respective correlation matrix (standardized PCA). The unstandardized method has the advantage of giving minimal weighting to the lower-order components (with smaller eigenvalues), which mostly define the noise element of the data, whereas the standardized PCA equally-weights all eigenvalues (this way inflating the weighting of variables with relatively small variance and reducing that of variables with greater variance) with the resulting exaggeration of the effects of noise (Mimmack *et al.* 2001). This second approach, however, is better able to separate the signal to noise ratio in the first components (Eklundh & Singh 1993). In this approach, by discarding the last components it is possible to eliminate most of the data noise. This data reduction procedure, on the other hand, is not recommended for subsequent use in a regression analysis, as the principal components with the smallest variation may describe infrequent thought influential events, which can be of greater importance in the regression equations (Jolliffe 1982). Hence, we analysed the 12 spectral bands (in a stack) and retrieved 12 principal components, as calculated by the unstandardized scores.

- *Model building*

For the purpose of defining the different sets of predictor variables, all products derived from the RS data were matched with the bird census data (within circular-plots of 125 m radius). This procedure was done by generating a dataset which synthesizes the predictor data within the circular-plot, to be used as model training data. The *Hard*, *Soft*, *SIs*, *Ord* and *Raw* datasets were calculated by averaging the image pixel values, weighted by their respective proportional composition in the circular-plots. This resulted in proportions of cover of each of the hard classes, and mean values for each of the other methods. At this stage we excluded the *Hard* and *Soft* classes *Woodland / Shrubs* (May), as we considered there not to have been any changes in the cover of woodlands and shrubs during the period of study. Indeed, these corresponded to roughly the same areas in the classified images and therefore the predictor related to the highest performing classifications (March) was selected. The class *Water* (for both March and May) was also removed as it was not considered relevant to the studied species.

The species-environmental data were fitted with MARS (Multivariate Adaptive Regression Splines) models (Friedman 1991), implemented in R statistical software (R Development Core Team 2008) with modified code from the *mda* package (Hastie & Tibshirani 1996), to allow for binary data (logit link function) and n-fold model cross-validation (Elith & Leathwick 2007). MARS is a nonparametric regression approach, which is capable of fitting complex (non-linear) responses with reduced computational resources, by applying piecewise linear regression functions in a (modified) recursive partitioning manner (Friedman 1991). Its is therefore a fast and high performing modelling method (De Veaux *et al.* 1993), having been successfully applied in SDM studies (Leathwick *et al.* 2005; Elith & Leathwick 2007). The models are built in a two-step approach: an initial forward stepwise procedure which iteratively splits the data domain into sub-regions (by placing knots) to fit them with a series of basis functions by ordinary least squares linear regression, until convergence; and a subsequent backward stepwise deletion strategy to produce an optimal set of basis functions, which keep most of the predictive power. This backward pass, aimed at eliminating model overfitting, is based in the lack-of-fit (LOF) criterion, a modification of the generalized cross-validation (GCV) criterion (Craven & Wahba 1979) to account for non-linearity, which penalizes both for lack-of-fit and increasing number of basis functions (Friedman & Silverman 1989; Friedman 1991), acting as a model regularization parameter. The implemented code allows the definition of the "penalty" parameter (default = 2), a cost function of the maximum number of terms (basis functions) included in the model, used to calculate the LOF criterion (Friedman 1991; Elith & Leathwick 2007). Other features of MARS include the capability of fitting interactions between predictors ("mars.degree" parameter in the code, which refers to the maximum interaction order and is by default set to 1, this way fitting additive models) and multiple-species responses (potentially good for rare species with few data records).

Owing to its forward selection procedure, MARS is vulnerable to high collinearity (or concurvity) in the predictor variables, with a potentially significant performance loss. Indeed, the choice of the best fitting variable (or knot) at a particular step of the procedure determines the choice of all further variables (and knots), and may not generate the best possible final variable combination (De Veaux & Ungar 1994). Additionally, high collinearity in the data presents severe problems for model interpretability (Friedman 1991; Morlini 2006). For these reasons we applied a variable reduction approach to each of the model training datasets so that they only incorporated predictor variables with pairwise correlation values smaller than 0.7 (Freedman *et al.* 1992). For this purpose, however, we used the Spearman rank correlation instead of the Pearson coefficient in order to account for some of the non-linear correlation effects. When two highly correlated candidate variables were found, we chose the one with the best average fit (as measured by the deviance explained) across species, in univariate MARS models.

Table 3.6 - Predictor variables used in the five training datasets (Legend: Mar – March; Dry veg. – Dry vegetation; Green veg. – Green vegetation; Wd/shrub – Woodland/shrubs; PC1 to PC12 – Principal Components 1 to 12; TM1 to TM5 and TM7 – Thematic Mapper spectral bands 1 to 5 and 7)

Hard	Soft	SIs	Ord	Raw
Bare soil (Mar)	Bare soil (Mar)	TC Bright (Mar)	PC1	TM1 (Mar)
Low veg. (Mar)	Cereal (Mar)	NDVI (Mar)	PC2	TM2 (Mar)
Fallow (Mar)	Bare soil (May)	TC Bright (May)	PC3	TM3 (Mar)
Cereal (Mar)	Fallow	NDSVI (May)	PC4	TM4 (Mar)
Bare soil (May)	Cereal (May)	TC Wet (May)	PC5	TM5 (Mar)
Low veg. (May)	Dry / Green veg. (May)		PC6	TM7 (Mar)
Fallow (May)	Woodland / Shrubs		PC7	TM1 (May)
Cereal (May)			PC8	TM2 (May)
Dry veg. (May)			PC9	TM3 (May)
Green veg. (May)			PC10	TM4 (May)
Woodland / Shrubs			PC11	TM5 (May)
			PC12	TM7 (May)

No significant correlations were found between variables in the *Hard* dataset so all were kept (Table 3.6). The *Soft* dataset included two pairs of highly correlated variables: *Fallow* (March) and *Fallow* (May) (Spearman *rho* $r_s = 0.799$; $n = 1100$; $p < 0.0001$); and *Dry vegetation* (May) and *Green vegetation* (May) ($r_s = -0.726$; $n = 1100$; $p < 0.0001$). The first pair was reduced to *Fallow* (May), which should hereafter be representative of the class *Fallow* throughout the period of study. The second pair was reduced to *Dry vegetation* (May), which should then represent the vegetation phenological gradient *Green / Dry vegetation.* The analysis of the *SIs* dataset aimed at eliminating highly correlated spectral indices but also, at a conceptual level, eliminating indices with similar interpretation. The inspection of the respective correlation matrix showed that, for each date (March and May), all green vegetation and soil moisture indices were highly and positively inter-correlated. The same was the case for all the previous and the NDSVI in March but not in May, probably due to the absence of senescent vegetation (measured by the NDSVI) at the earlier date. Additionally, we found that the NDSI was always highly and negatively correlated with all green vegetation and moisture indices, within each date. After applying the defined variable reduction approach the dataset was reduced to five predictors which are able to describe the spatio-temporal patterns of the main relevant landscape features: bare soil, green and senescent vegetation and soil moisture (Table 3.6). The *Ord* dataset showed no significant correlations between variables, confirming that the PCA was capable of breaking down not only the data's linear dependencies but also its rank correlation patterns. The *Raw* dataset, which is to be considered as control, was not subject to this variable reduction procedure. This dataset illustrates the potential of raw (unprocessed) RS data for direct use as predictors in the SDMs.

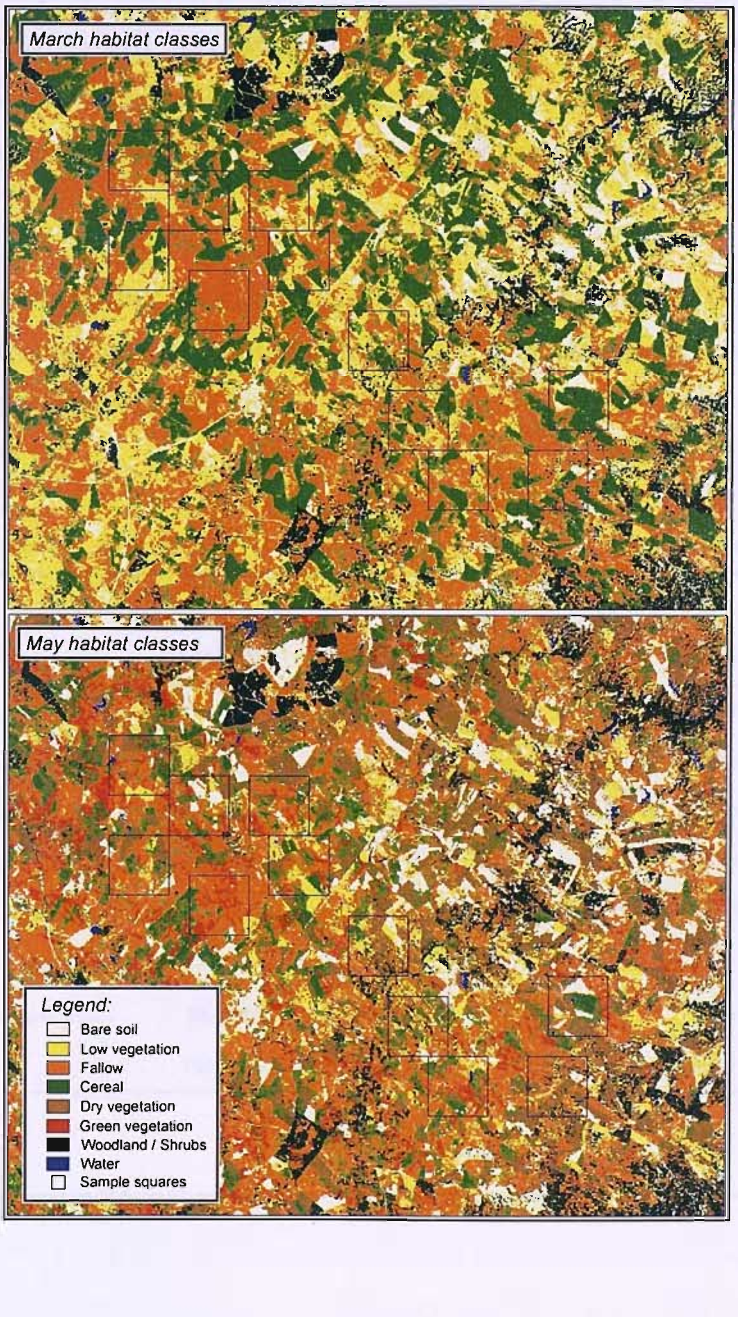
Single-species MARS models were fitted on each of the five datasets. The models' maximum interaction order and penalty were optimised according to the best average model performance for all methods and all species. This generated a common methodological approach, which allowed an easier comparison of the model results. Model performance was assessed through a 10-fold cross-validation, while controlling for prevalence in the data resampling. In order to account for the variability inherent in the cross-validation process (Breiman 1996), five replications of each cross-validation were ran and the respective average calculated. As a performance measure we calculated the respective averaged ROC AUC scores (Hanley & McNeil 1982). Additionally, the same measure was calculated directly on the training data, which, through comparison with the cross-validated value allows inference about the model generality.

3.3. Results

- *Digital image Processing*

The image hard classification approach generated two classified maps corresponding to the habitat-related classes for March and May dates (Figure 3.2).

Figure 3.2 - March (above) and May (below) habitat maps, as resulting from the image hard classification, overlaid with the sample squares



Both classifications achieved high overall performance as calculated by the cross-validation procedure with OCA values of 90.67% and 87.80%, and Kappa of 0.8589 and 0.8380, respectively for the March and May habitat classes (Table 3.7). The respective user's accuracies ranged from 87.03% and 100.00% (92.10% on average) and from 80.07% and 100.00% (89.87% on average), and the producer's accuracies from 64.71% and 94.65% (83.92% on average) and from 70.59% and 92.61% (84.05% on average). The lowest of these values (64.71% and 70.59%) are relative to the producer's accuracies for class *Woodland / Shrubs* on both March and May classifications. This measure refers, however, to the proportion of pixels of a certain class (according to the ground truth information) which were correctly classified. Since this class is not well represented in the validation dataset, the accuracies obtained should be considered an underestimation of the real classwise performances.

The average performance increase resulting from the incorporation of the full (multi-date) spectral data on both March and May habitats classification was of 8.07% in OCA and 12.77% in Kappa (Table 3.8).

Table 3.7 - Performance of the image hard classification into the March and May habitat-related classes

	March		***May***	
Class	***User's Acc.***	***Prod.'s Acc.***	***User's Acc.***	***Prod.'s Acc.***
Bare soil	92.74 %	86.47 %	93.68 %	92.61 %
Low vegetation	87.03 %	82.08 %	80.07 %	75.79 %
Fallow	89.22 %	94.65 %	88.89 %	92.47 %
Cereal	95.62 %	92.29 %	86.00 %	80.48 %
Dry vegetation	-	-	87.36 %	89.22 %
Green vegetation	-	-	90.67 %	87.93 %
Woodland/Shrubs	88.00 %	64.71 %	92.31 %	70.59 %
Water	100.00 %	83.33 %	100.00 %	83.33 %
OCA	90.6699 %		87.7990 %	
Kappa	0.858917		0.838034	

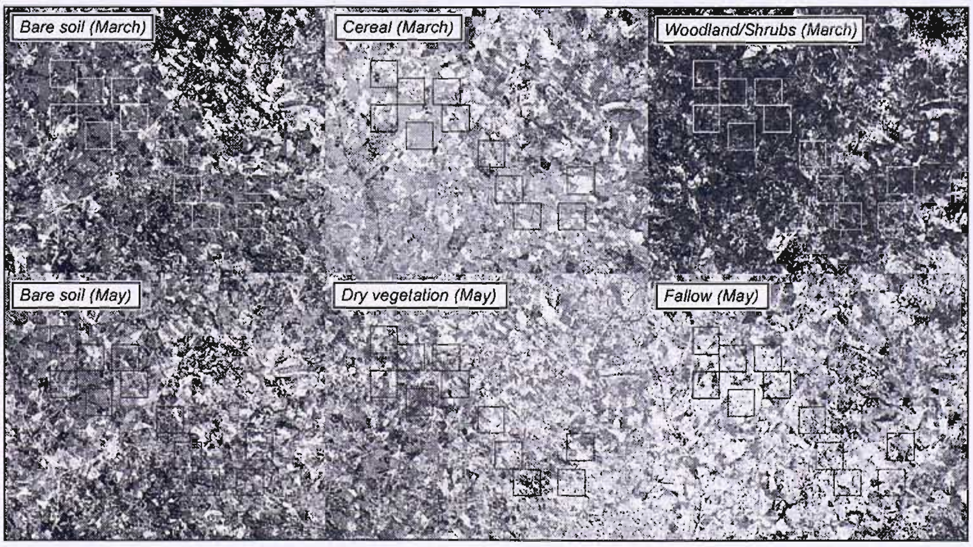
The fact that the additional data (from the classes *Woodland / Shrubs* and *Water*) was not incorporated in the validation dataset did not allow for a statistical validation of its inclusion in the image classifications. It was, however, possible to observe improvements in the discrimination between these two classes, by visual inspection of the classified imagery in areas of known cover.

Table 3.8 - Performance comparison between the use of single-date and multi-date spectral data in the image hard classification models

	Single-date spectral data	Multi-date spectral data
March classes		
OCA	85.6459 %	90.6699 %
Kappa	0.781883	0.858917
May classes		
OCA	79.4942 %	87.7990 %
Kappa	0.722878	0.838034

The image soft classifications generated a probability of membership map for each class, on each classification (Figure 3.3).

Figure 3.3 - Examples of soft classification outputs (the gradient black to white reflects the probability of membership of a pixel to each class, ranging respectively from 0 to 1): Bare soil (March); Cereal (March); Woodland / Shrubs (March); Bare soil (May); Fallow (May); and Dry vegetation (May)



The overall cross-validated performance values achieved by these models were OCA of 95.53% and 91.18% and Kappa of 0.9171 and 0.8754, respectively (Table 3.9). The user's accuracies achieved ranged between 95.29% and 100.00% (96.97% on average) and between 86.52% and 100.00% (92.19% on average), and the producer's accuracies between 76.47% and 97.74% (91.21% on average) and between 71.43% and 96.02% (86.70% on average).

Table 3.9 - Performance of the image soft classification into the March and May habitat-related classes

	March		***May***	
Class	***User's Acc.***	***Prod.'s Acc.***	***User's Acc.***	***Prod.'s Acc.***
Bare soil	97.74 %	97.74 %	95.48 %	96.02 %
Fallow	95.29 %	97.62 %	92.38 %	94.52 %
Cereal	95.51 %	92.54 %	86.52 %	82.35 %
Dry vegetation	-	-	90.16 %	91.34 %
Green vegetation	-	-	91.48 %	87.93 %
Woodland/Shrubs	96.30 %	76.47 %	89.29 %	71.43 %
Water	100.00 %	91.67 %	100.00 %	83.33 %
O.C.A.	95.5346 %		91.1844 %	
Kappa	0.917136		0.875363	

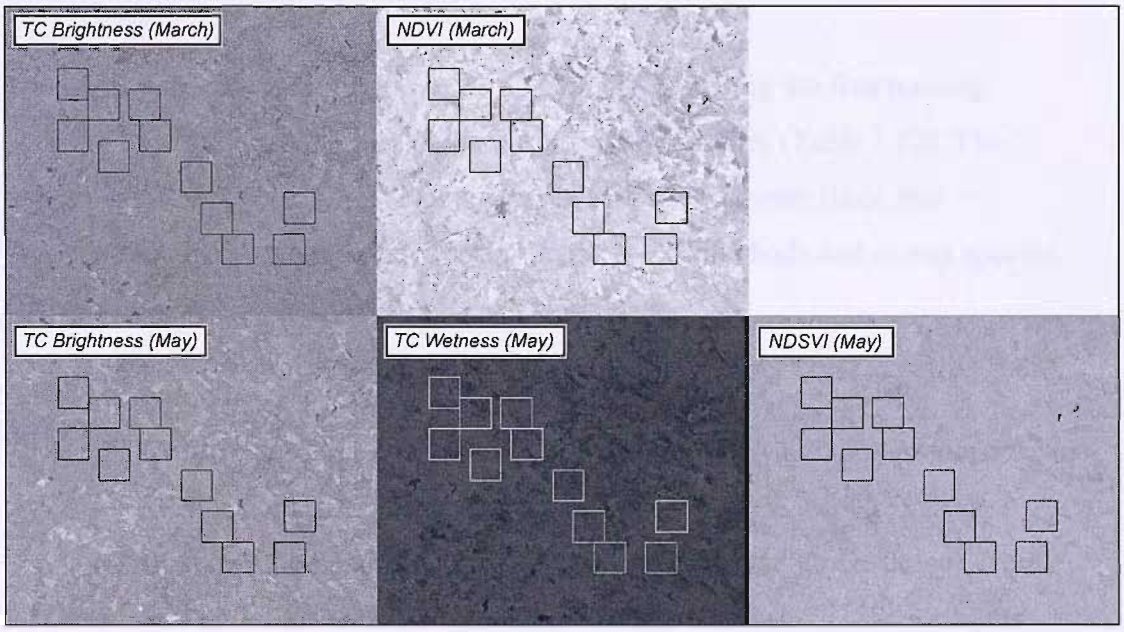
The resulting performance increase due to incorporation of the multi-date spectral data was, on average, 9.05% in OCA and 15.74% in Kappa for both classifications (Table 3.10).

The spectral index calculation process resulted initially in 18 images (9 indices x 2 dates), which were further reduced to 5 by the variable reduction procedure above explained (Figure 3.4).

Table 3.10 - Performance comparison between the use of single-date and multi-date spectral data in the image soft classification models

	Single-date spectral data	Multi-date spectral data
March classes		
OCA	92.2284 %	95.5346 %
Kappa	0.854328	0.917136
May classes		
OCA	78.9894 %	91.1844 %
Kappa	0.694354	0.875363

Figure 3.4 - Spectral indices used in the models: TC Brightness (March); NDVI (Mar); TC Brightness (May); TC Wetness (May); and NDSVI (May)



By the band ordination method we obtained 12 principal components. Most of the data variability was contained in the higher order components: 95.87 % in the first 4 components (Table 3.11).

Table 3.11 - Percentage variability contained in each of the 12 principal components

Principal Component	Variability	Principal Component	Variability
1	47.015 %	7	0.555 %
2	29.360 %	8	0.314 %
3	12.650 %	9	0.190 %
4	6.842 %	10	0.091 %
5	1.923 %	11	0.069 %
6	0.937 %	12	0.053 %

- *Species distribution models*

The MARS models were fitted for the 13 bird species using the five training datasets, corresponding to the five data analysis approaches (Table 3.12). The default values of "penalty" (= 2) and "mars.degree" (= 1) were those that presented the best average model performance for all methods and across species, so these parameters were fixed at these values.

The model performance values, as calculated by the cross-validation procedure (AUCcv), ranged respectively from 0.518 to 0.819 (0.683 on average), from 0.595 to 0.814 (0.689 on average), from 0.542 to 0.823 (0.683 on average), from 0.569 to 0.828 (0.710 on average) and from 0.576 to 0.844 (0.703 on average), on datasets *Hard*, *Soft*, *SIs*, *Ord* and *Raw*. The mean difference found between these measures and the AUC calculated on the training data was of 0.082, 0.083, 0.072, 0.088 and 0.099 for the respective datasets.

Overall, the numerical-based methods achieved higher performances than the knowledge-based approaches (Figure 3.5). Additionally, the use of dataset *SIs*, resulted in good models, i.e. with AUCcv value above 0.7 (Hosmer & Lemeshow 2000) for five of the studied species and datasets *Hard* and *Soft*, for six species.

Models with datasets *Ord* and *Raw* achieved good performing models for seven and eight species, respectively.

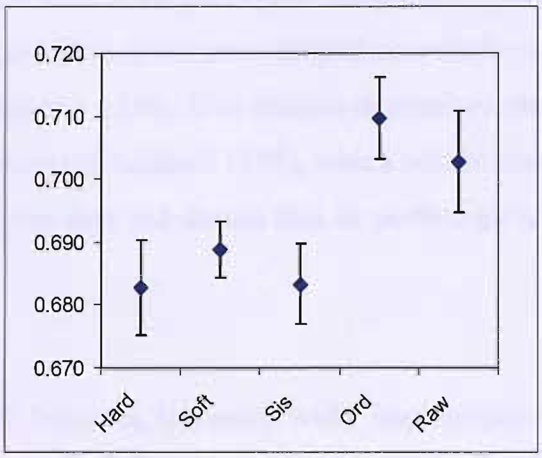
Table 3.12 - Model performance for the 5 datasets, as calculated directly on the training data (AUC) and through a 10-fold cross-validation (AUCcv)

	Hard		***Soft***		***SIs***		***Ord***		***Raw***	
Species	**AUC**	**AUCcv**	**AUC**	**AUCcv**	**AUC**	**AUCcv**	**AUC**	**AUCcv**	**AUC**	**AUCcv**
Milcal	0.804	0.766	0.770	0.727	0.785	0.768	0.828	0.791	0.810	0.761
Melcal	0.781	0.752	0.752	0.711	0.768	0.745	0.824	0.788	0.830	0.782
Galsp	0.787	0.756	0.774	0.741	0.732	0.702	0.824	0.783	0.821	0.773
Tettet	0.716	0.644	0.727	0.673	0.697	0.645	0.771	0.722	0.767	0.712
Cisjun	0.858	0.819	0.852	0.814	0.852	0.823	0.879	0.828	0.883	0.844
Saxtor	0.738	0.681	0.726	0.654	0.715	0.660±	0.774	0.721	0.785	0.724
Calbra	0.782	0.700	0.778	0.739	0.757	0.727	0.819	0.735	0.794	0.734
Aleruf	0.764	0.685	0.742	0.650	0.708	0.675	0.764	0.669	0.788	0.702
Cirpyg	0.734	0.621	0.749	0.657	0.699	0.607	0.772	0.646	0.701	0.662
Antcam	0.834	0.685	0.836	0.705	0.812	0.696	0.804	0.666	0.845	0.591
Otitar	0.650	0.518	0.778	0.617	0.732	0.623	0.766	0.644	0.827	0.584
Oenhis	0.796	0.705	0.747	0.671	0.840	0.671	0.833	0.662	0.862	0.686
Buroed	0.698	0.544	0.796	0.595	0.722	0.542	0.705	0.569	0.709	0.576

3.4. Discussion

This chapter started by reviewing some of the most commonly used RS data sources, their characteristics and main areas of application. The large number of available sensors, collecting information at a wide range of spatial scales, makes them a prime source of environmental descriptor data, which can be used as predictor variables in SDMs. Nevertheless, the empirical nature of these models makes them particularly sensitive to errors or biases in the input data (Buckland & Elston 1993; Elith *et al.* 2002). These biases can be within the species locational dataset, like observer bias (Diefenbach *et al.* 2003), sampling bias (see Chapter 2) and positional errors (Osborne & Leitão *in press*) (see Appendix A.2), or as well in the model predictors. In order to reduce the possible sources of bias in RS data, all the steps taken in the process of information extraction (image pre-processing and processing) should be chosen with care.

Figure 3.5 - Mean (and standard error) of the model performance values, across species, for the 5 datasets



The first possible source of bias is the selection of the RS data source itself. This requires a good understanding of the data characteristics of the different sensors, their spectral, radiometric, spatial and temporal resolutions (Lu & Weng 2007). The selected imagery should be able to characterise the features or habitats that determine the species distributions, at an appropriate spatial resolution (Kerr & Ostrovsky 2003; Guisan *et al.* 2007). Image availability and cost, as well as atmospheric conditions can also be influential for this choice. For example, in regions of high cloud cover persistence (as in some tropical areas) it is difficult to obtain a good, cloud-free, ground coverage from optical sensors and thus radar systems could be preferable. In our case study, the Landsat TM imagery acquired was able to describe the landscape vegetation patterns important for the studied species. Additionally, by acquiring two overlapping images from along the season, we were able to describe the observed phenological events in the area, during the period of study.

The image pre-processing stage includes the radiometric, atmospheric and geometric corrections of the images. The radiometric and atmospheric corrections directly change the values of the image pixels, reducing the biases from the sensor (converting them to radiance or reflectance values) and from atmospheric effects, such as haze scattering. Moreover, when using multi-date (or multi-sensor) imagery, these corrections allow for a direct comparison between different imagery and should therefore be mandatory (Paolini *et al.* 2006). Geometric correction is the process of image resampling (by interpolation) and registration to the user-defined coordinate system, this way allowing data integration in a GIS. This process determines the final image pixel values and spatial patterns (Cracknell 1998), which will be used as environmental descriptor data and should thus be performed with the best possible accuracy.

The DIP method used, however, has much wider implications for the subsequent analysis, as it defines the model predictor variables, and thus influences model performance, interpretability and generality. Therefore, the selection of an adequate approach should be carefully assessed. Our study aimed at providing a comparison between different possible methods for information extraction of these data.

Image classification approaches, by definition, bias the spectral information by assigning it to spectral or thematic classes, through the definition of thresholding rules. Unsupervised methods, although requiring little input (bias) from the user (such as number of output classes or number of iterations to use in the calculations) and thus being closer to the spectral data distribution, may reveal classes that are not important for describing the species distribution patterns. Supervised techniques, on the other hand, can directly relate to the desired habitat classes, which should be defined accordingly. The classification system should thus be designed in order to suit the study aim, by incorporating existing knowledge of the species ecology. The prior definition of the output classes, however, is a source of data bias and should be carefully undertaken. The chosen classes should be suitable to describe the species distribution patterns, on which depends the good performance of the SDMs. Nevertheless, the definition of non-interesting classes can be useful in the image classification process, as it may improve the spectral separability of the classes of interest. A sufficient number of training samples should be selected, capable of a good spectral characterisation of all defined classes. Indeed we found that the classification of the habitat classes in the Castro Verde study area benefitted from the inclusion of additional training points of two classes which were poorly represented in the study area (*Woodland / Shrubs* and *Water*). Additionally, and depending on the classification method used, the inclusion of pixels of mixed cover in the training data can further bias the class spectral characterisation.

Image hard classification, i.e. the assignation of image pixels into discrete classes, is the most commonly used DIP approach. Parametric methods for image classification, such as the MLC, can work well to distinguish homogeneous classes and with good spectral separability. However, they assume a normal distribution of the input (spectral) data that usually would require an additional processing step (for data normalisation). In addition, these methods fail to describe classes with complex or overlapping spectral signatures. Thus, non-parametric approaches such as ANN or SVM are desirable in order to achieve good classification results (Benediktsson *et al.* 1990; Huang *et al.* 2002a). These more robust methods are also capable of dealing with small or mixed pixels training datasets (Foody & Mathur 2006). We aimed to achieve a classification system which was ecologically meaningful, even though with the definition of thematic classes with complex and intergrading spectra. Nevertheless, by carefully defining a classification system that incorporated a good knowledge of the area and the respective landscape dynamics, it was possible to achieve good overall and classwise performances.

Soft classification techniques, by keeping quantitative information relative to the probability of membership of each pixel to every class, include one less step of information reduction in the process and should therefore be less biasing than hard classification approaches. On the other hand, by imposing a classification system to the data they allow an interpretation in terms of the classes of interest in the specific study. Moreover, they are able to describe the vegetation continua and intergrading classes as they occur in natural or semi-natural landscapes (Foody 1992). They present, therefore, a compromise between information loss (and bias) and class interpretability, which should be ideal for use in ecological studies. Once again, the choice of the classification algorithm can influence the final classification performance. The classification accuracy assessment was, however, based on measures most adequate for hard classification approaches (Gómez *et al.* 2008), which should nevertheless be good indicators of the soft classification results.

The use of spectral indices aims to extract useful information from specific spectral band combinations known to relate to particular surface characteristics, such as vegetation vigour, soil moisture or amount of bare soil. These indices, by making use of the information contained in several spectral bands into single meaningful measures, can be considered as data reduction techniques. Different indices reduce the data in different ways, and the choice of which indices to use is therefore another source of bias. Most spectral indices, in fact, combine only two spectral bands and discard most of the data contained in the original spectral bands. The Tasselled Cap indices (Kauth & Thomas 1976; Crist & Cicone 1984), on the other hand, make use of the six reflective bands of the Landsat TM sensor. However, the index selection should be related to the aim of the study, by being able to describe the species occurrence patterns and by providing good ecological interpretability. For this reason we made our decision depend on the species data, by fitting univariate models and selecting those indices that better described (on average, across species) the observed patterns. This approach is not new (Zimmermann *et al.* 2007) and aims to find a best set of common predictor variables for all species, while avoiding collinearity problems in the modelling process. Indeed, the Tasselled Cap indices were preferred in most cases to the simpler indices, with the exception of the NDVI for March, which was preferred instead of the TC Greenness index. The presence of senescent vegetation in May, and its relevance for the studied species, resulted in the selection of the NDSVI for this date. The five indices selected by this approach were then able to describe the amount of bare soil, green vegetation (and soil moisture, which were always highly correlated) and senescent vegetation in both dates.

The three previous approaches (hard and soft image classifications and spectral index calculation) generate model predictor variables which are ecologically interpretable, with varying degrees of data reduction and bias introduced. The different habitat classes described by the image classifications should be considered indirect predictors, as they indirectly reflect a combination of the resources used by the species (Guisan & Zimmermann 2000). The nature of the predictors extracted from the two classification methods is therefore identical, even though the former implies more information loss by the assignation of each pixel to individual classes. On the other hand, the latter may be more difficult to interpret, as the proportion of cover of a certain class is more parsimonious than the probability of membership to the same class. Spectral indices, in turn, can express resource, direct and indirect predictors. Amount of green vegetation, as measured by the NDVI could be a resource for some of the species (being directly consumed by them), or an indirect predictor for others. The amount of bare soil present (TC Brightness index) reflects an indirect predictor, the same way as the bare soil class from the classification approaches. Surface soil moisture or humidity, as measured by the TC Wetness index, can be considered a direct predictor as it has a direct physiological influence on the species. Thus, the image classification approaches generate more distal predictors than the spectral indices (which are more proximal), and should result in models with less generality than the latter method. However, by being easily translatable into land use practices, they may have an additional value for application to local conservation actions.

The ordination of the spectral bands into principal components for use in regression models is a classical approach to handle data collinearity, known as principal component regression (Hotelling 1957; Massy 1965), which in the case of MARS models has been referred to as principal component MARS (De Veaux & Ungar 1994). Indeed, the PCA applied to our spectral dataset was able to break the linear and the rank correlation effects between bands. However, the MARS models, by fitting non-linear responses to the data, are sensitive to data concurvity (non-linear dependencies) (Friedman 1991), hence the implementation of newer non-linear transformation approaches, such as the Additive PCA (Donnell *et al.* 1994) or the iterative Kernel PCA (Schölkopf *et al.* 1997; Kim *et al.* 2005) would have been preferred to the intended end. Nevertheless, these ordination approaches aim to keep all the original data variability in a new (non-correlated) feature space, for use in the regression models. The same way, the direct use of the raw spectral bands in the models also keeps all the data information, even though in a form that is not fully usable by the models, due to data correlation effects, as discussed before. Consequently, the models resulting from this approach have lower performances than those of the principal component MARS. These two numerical-based approaches, while keeping most of the data variability and this way resulting in better performing models, result in distal and indirect variables of difficult (if possible) interpretation.

Our results confirmed the higher model performance (on average) of the numerical-based approaches in relation to the knowledge-based ones. This finding suggests that these methods are preferable when issues of model generality or interpretability are not concerned, i.e. when used only for prediction within the study site and period. This is the case when the aim of the study is to derive the best possible distribution map of the studied species, without requiring an ecological interpretation or model projection to other areas or study period. Nevertheless, a post-modelling interpretation of the resulting patterns and association with ecological meaningful variables is always possible. From these two approaches, and due to the modelling process itself, we found that data ordination resulted in best performing models, whereas the raw bands produced models with a higher degree of overfitting.

When an ecological interpretation is to be taken from the model results, knowledge-based approaches must be used. From the comparison between these methods we found that predictors generated by image soft classification resulted in better performing models (on average) than those of the two other approaches. This highlights the potential of soft classification methods for use in SDMs.

Models fitted on spectral indices, however, should have more generality than those on habitat classes, due to the more proximal nature of their predictors.

At an individual species level, however, it was not possible to conclude on a best predictor-deriving procedure. Within the knowledge-based approaches, both hard class and soft class predictors resulted in the best performing model for five species each (*Melcal*, *Galsp*, *Saxtor*, *Aleruf* and *Oenhis* with hard classes and *Tettet*, *Calbra*, *Cirpyg*, *Antcam* and *Buroed* when using soft classes). The use of spectral indices as predictors resulted in the best models for three species (*Milcal*, *Cisjun* and *Otitar*). Moreover, even though the numerical-based methods generally resulted in better performing models, this was not always the case, and some species (*Calbra*, *Antcam*, *Oenhis* and *Buroed*) were better modelled by knowledge-based predictors. Lastly, it is expected that these models could be further improved with the incorporation of predictor variables from other sources, such as those describing topography (slope, roughness, etc) and landscape or disturbance features, such as distance to road, to track, to water or to tree, etc.

- *Conclusions*

Remote sensing data, from the multiple existing sources and using the varied image processing methods available, provide a wide range of environmental descriptors, which can be used as predictor variables in SDMs. However, the choice of imagery to use and of the processing method to apply, determine the characteristics of these predictors, with a resulting effect on model performance, their generality and interpretability. Most importantly, DIP methods dictate the type and amount of information retained in the final predictors. The most data reductive methods, either by the definition of thresholds, such as image classification methods, or by discarding information, like the calculation of spectral indices, generally result in poorer performing models as judged by the cross-validated AUC values. On the other hand, numerically-based methods which keep the full data variability, such as PCA or even the direct use of the raw spectral bands as predictors, although resulting in higher performing models lack ecological interpretability. The selection of the DIP method to use in the RS imagery should therefore be undertaken with care, always accounting for the advantages and limitations of the respective method and its adequacy to the study aims.

4. Very-high resolution laser altimetry data for describing landscape features in the Castro Verde study area: the STEPPEBIRD campaign

4.1. Background

The Castro Verde study area, being the main steppe bird area in Portugal, holds populations of several threatened steppe bird species with national and international importance (Moreira *et al.* 2007) (see Appendix A.1). Despite the importance of cereal steppes to conservation, they have changed significantly during recent decades through agricultural intensification, land abandonment and afforestation with direct impact on bird populations (Baldock 1991; Tucker & Heath 1994; Suárez *et al.* 1997). In order to devise adequate management schemes for the conservation of steppe birds and their habitats, it is necessary to understand how species use the environment. This, however, requires the acquisition and extraction of adequate environmental descriptors, which relate to the observed species occurrence patterns. These descriptors, when used in species distribution models (SDMs) allow the inference of species habitat preferences and the prediction of their distributions (Guisan & Zimmermann 2000).

Besides habitat type (explored in Chapter 3), vegetation height has been previously referred to as a key aspect that could influence bird populations in these landscapes (Moreira 1999). The presence or density of trees, by acting as landscape fragmenting features, has also been found to influence some of these species (Moreira *et al.* 2007; Reino *et al. in press*) (see Appendix A.1). Also, built-up structures (like farmhouses) can be interpreted as indicators of disturbance and potentially be used to derive a predictor of species distributions. Finally, some studies at a larger scale have found topographic variability to influence some of these species as they tend to prefer relatively flat areas (Osborne *et al.* 2001; Suárez-Seoane *et al.* 2002a).

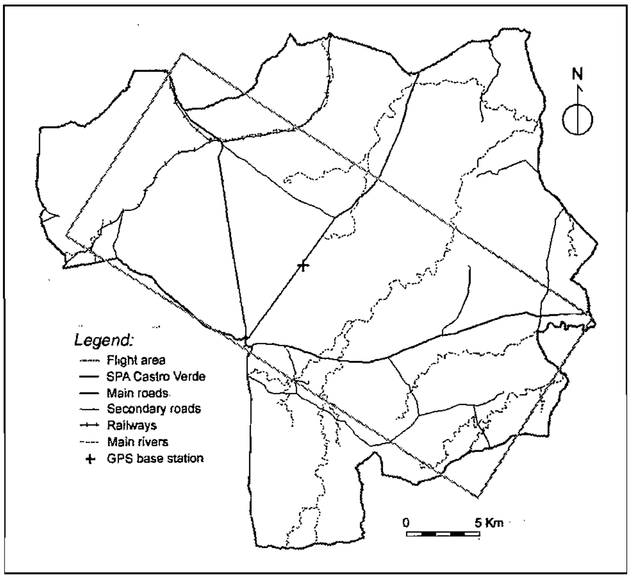
Airborne laser altimeter data, also known as Light Detection And Ranging (LiDAR) data, are capable of quantifying topography and vegetation canopy properties (Ritchie 1996). More specifically, LiDAR data are capable of measuring vegetation height (Davenport *et al.* 2000; Genç *et al.* 2004) and the underlying topographic height (Cobby *et al.* 2001). Other landscape features, such as farmhouses and other built-up structures, can also be described by these data (Maas & Vosselman 1999; Priestnall *et al.* 2000). Moreover, these data, when coupled with high-resolution multispectral imagery may be used to derive spatially fine-grained predictors over large areas (Mason *et al.* 2003; Bradbury *et al.* 2005).

The current chapter aims to describe the EUFAR (European Fleet for Airborne Research) STEPPEBIRD research project, including a flight taken over the Castro Verde study area in the Spring of 2006, the data collected by the sensors on-board the aircraft and in the field campaigns, as well as the data processing for feature extraction and the generation of environmental descriptors for use as predictors in SDMs within the study area.

4.2. Air campaign

A Dornier 228-101 aircraft, from the Natural Environmental Research Council (NERC) Airborne Research & Survey Facility (ARSF), flew over the study area on the 18th and 19th of May 2006, as part of the "Western Mediterranean Campaign", and supported by EUFAR. On board, LiDAR data were collected by the *Optech Airborne Laser Terrain Mapper (ALTM) 3033* sensor. The flight navigation system used a DGPS location system, referenced to a GPS base station mounted on the "Pereiros" trig-point (roughly in the centre of the study area), with *a Leica System 1230* real-time GPS receiver provided by the NERC Geophysical Equipment Facility (GEF; Figure 4.1). Owing to sensor failure, only part of the study area was covered on the first flight day (seven flight lines). The remaining of the area was covered (five additional flight lines) on the second day of flight. On both days the aircraft flew at an approximate altitude of 2000 m, which resulted in a point data cloud with an approximate density of one point at every 1.9 m^2. For each laser pulse, first and last return elevation and signal intensity were collected. Additionally, co-registered multi-spectral data were collected by *the Itres Compact Airborne Spectographic Imager (Casi-2)* sensor, which is still under processing (Leitão *et al.* 2007) and hence it is not described here.

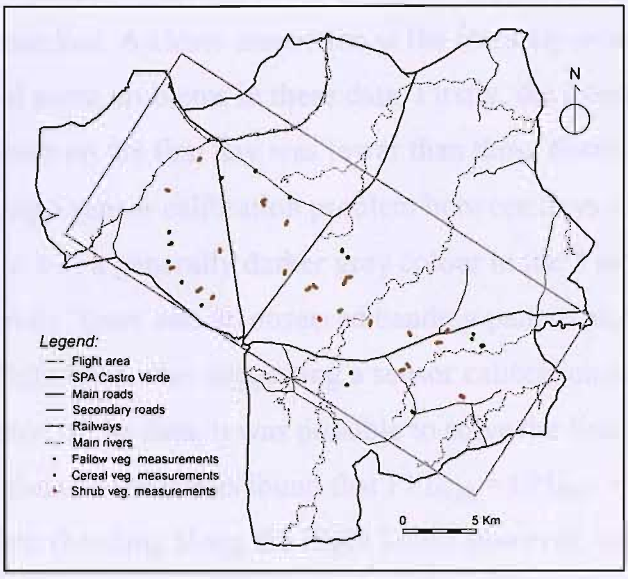
Figure 4.1 - Area covered by the STEPPEBIRD flight, and location of the GPS base station within the Castro Verde SPA



4.3. *Field campaign*

With the aim of calibrating the LiDAR vegetation height data, a series of vegetation height measurements were taken in the study area during the weeks around (before and after) the STEPPEBIRD flight (Figure 4.2). These measurements were made along transects defined according to the different relevant land uses (fallow, cereal and shrub), and evenly distributed throughout the study area. DGPS locations were recorded for all measurement samples. Fallow field measurements were done by sampling the whole spectrum of vegetation heights, from the intensively pastured fields (with very low vegetation) to the old full growth fallows (vegetation higher than 50 cm). In total, 401 measurements were made on fallow fields distributed along 19 transects. Fallow vegetation is often very heterogeneous in height, having more than one vertical stratum. Thus, sample points were selected along the defined transects, always spaced at least 5 m apart, so that they were evenly structured in a 3 m radius circle, and the dominant stratum was measured and the respective proportion of ground cover estimated. The different cereal crops (varieties of wheat, oats and barley) were measured by sampling the full range of observed heights. In total 148 cereal measurements were taken along 18 transects. Additionally, 40 measurements were done on shrub areas.

Figure 4.2 - Location of the (fallow, cereal and shrub) vegetation measurements within the Castro Verde SPA and the STEPPEBIRD flight area



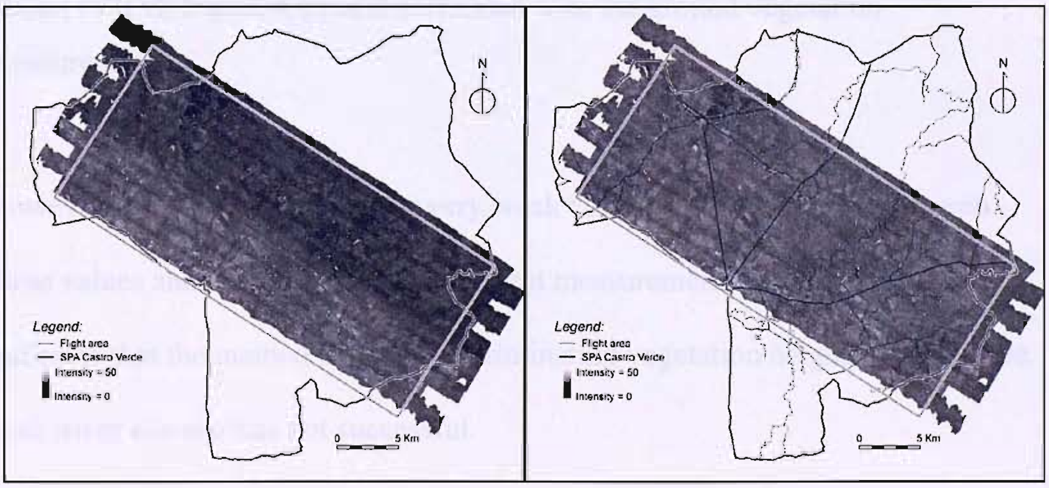
4.4. *Data processing and feature extraction*

The LiDAR data were originally processed by the Unit for Landscape Modelling of Cambridge University, as part of a collaborative arrangement with NERC ARSF. After this initial processing, these data were provided as point cloud data, in a single ASCII file per flight line, containing information on First Pulse Height (FPH), First Pulse Intensity (FPI), Last Pulse Height (LPH) and Last Pulse Intensity (LPI), geo-referenced to the UTM WGS84 coordinate system.

These data were initially geo-referenced to the "Gauss Militar" (Hayford-Gauss projection, International Ellipsoid, Datum Lisboa IGeoE) reference system, with the Azimuth Systems' *Azgcorr* software (Release 110). The four signals (FPH, FPI, LPH and LPI) were then split into individual point cloud data files for subsequent interpolation into raster images with a 5x5 m^2 pixel resolution. All data interpolations were done using the *GEON points2grid Utility* software (http://lidar.asu.edu/points2grid.html), with a search radius equivalent to half of the output pixel diagonal. After gridding (interpolating), the output images were mosaicked together by averaging the overlapping pixel values.

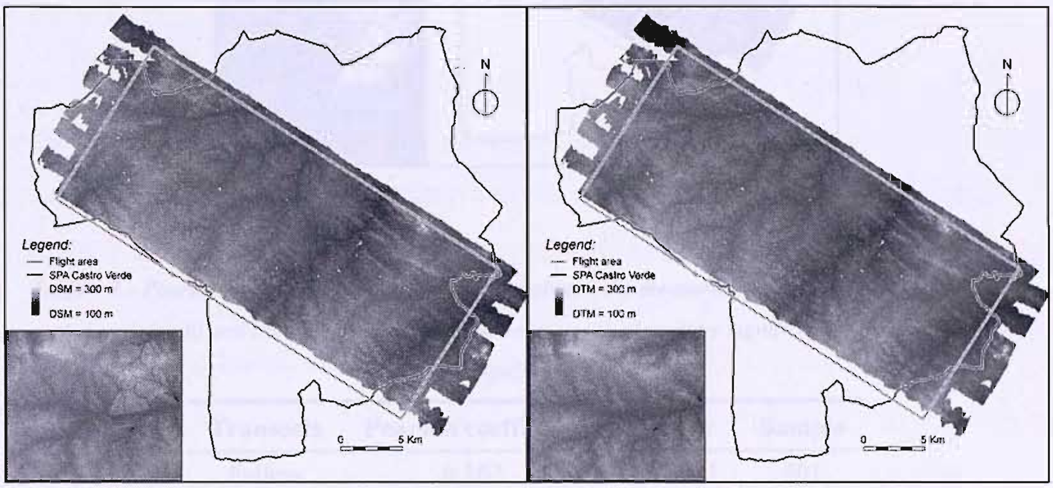
The FPI point cloud data were gridded by assigning the pixels to the mean elevation value of the points within the respective search radius, and subsequently mosaicked. A closer inspection at the intensity mosaic image, however, revealed some problems in these data. Firstly, the overall intensity in the flight lines flown on the first day was lower than those flown on the second flight day, denoting a sensor calibration problem between days – it is possible to observe (in Figure 4.3) a generally darker grey colour in the 7 most northern flight lines. Secondly, there was an observed banding pattern along and towards the edge of the flight lines, also suggesting a sensor calibration problem. Through empirical inspection of the data, it was possible to solve the first problem (between day calibration) as it was found that $FPI_{day2} = FPI_{day1} + 6.5$. The second calibration problem (banding along the flight lines), however, was not solved which makes the intensity data unusable without further processing.

Figure 4.3 - Mosaic of the LiDAR intensity signal before (left) and after (right) in-between day calibration correction



The FPH data was gridded by assigning the pixels to the maximum elevation of the points within the search radius. Assuming these data will be reflected from existing landscape features (like vegetation or built-up structures) this procedure generates a Digital Surface Model (DSM). It was also assumed that within each gridded pixel, at least some of the returned (first) pulse data was reflected from the ground (Streutker & Glenn 2006). Therefore the LPH data was gridded by assigning the pixels to the minimum elevation of the points within the search radius, this way generating a Digital Terrain Model (DTM). (Figure 4.4)

Figure 4.4 - LiDAR-derived DSM (left) and DTM (right), with detailed views (insets)



The difference between these two layers (DSM - DTM) was computed, which (in regions without built-up structures) should be equivalent to a Vegetation Height Model (VHM; Figure 4.5), and correlated with the ground vegetation measurements.

However, no significant and only very weak correlations were found between these values and the field vegetation height measurements (Table 4.1), which indicates that the method used for discriminating vegetation heights within these land cover classes was not successful.

Figure 4.5 - LiDAR-derived VHM, with detailed view (inset)

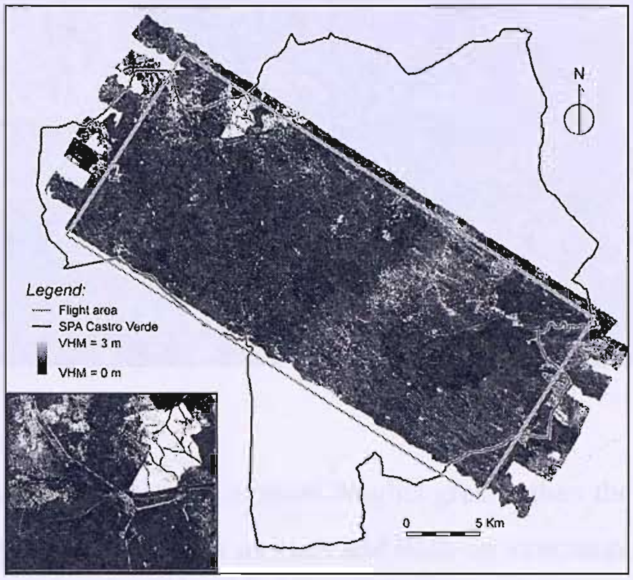
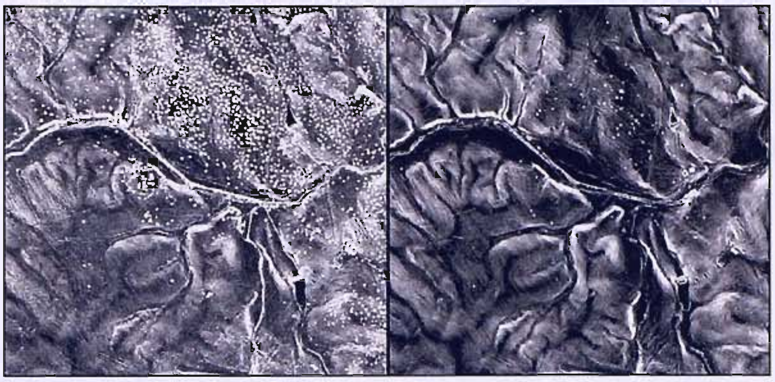


Table 4.1 - Pearson correlation coefficient values between the measured vegetation heights (in fallow, cereal and shrub) and the LiDAR-derived VHM values, with respective significance value (n.s. = non-significant)

Transects	Pearson coefficient	p-value	Sample
Fallow	0.162	< 0.0001	401
Cereal	0.075	n.s.	148
Shrub	-0.213	n.s.	40

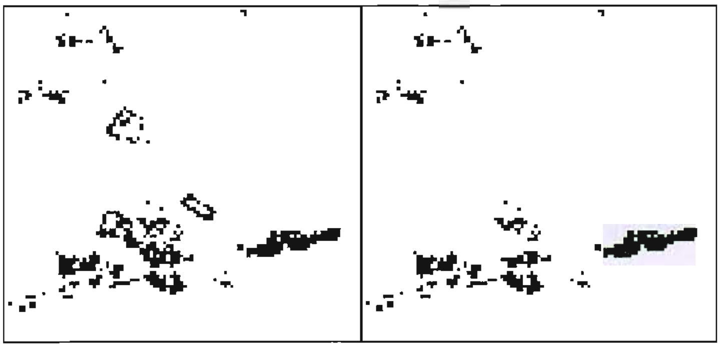
Instead, a high correlation between these and the surface slope (as calculated from the DTM) was found (r = 0.844; n = 589; p < 0.0001). Also by visual inspection of these two layers (VHM and Slope) it is possible to observe that the former mostly reflects the patterns of the latter with other features (such as trees) overlaid on top (Figure 4.6).

Figure 4.6 - Detailed view of the VHM (left) compared with Slope (right)



On the other hand, features with vertical heights greater than the slope range (within each gridded pixel), such as trees and built-up structures, should still be possible to extract from this VHM. Hence, a new image was generated which included all features with a height value above 3 m (in the VHM), which is then a map of (tall) vertical structures in the area (Figure 4.7). Both trees and built-up structures were identified in the resulting image, with the aid of topographic maps, the unprocessed CASI data and the Landsat soft classification (see Chapter 3) for the study region. A map of trees was obtained by cleaning the vertical structures map from built-up structures and a few particularly high cereal fields. Additionally, a map of built-up structures was generated by inclusion of all identified objects.

Figure 4.7 - Detailed view of the map of vertical structures above 3 m height (left, in black), compared with the tree map (right, in green), resulted from the identification and "cleaning" of the built-up structures

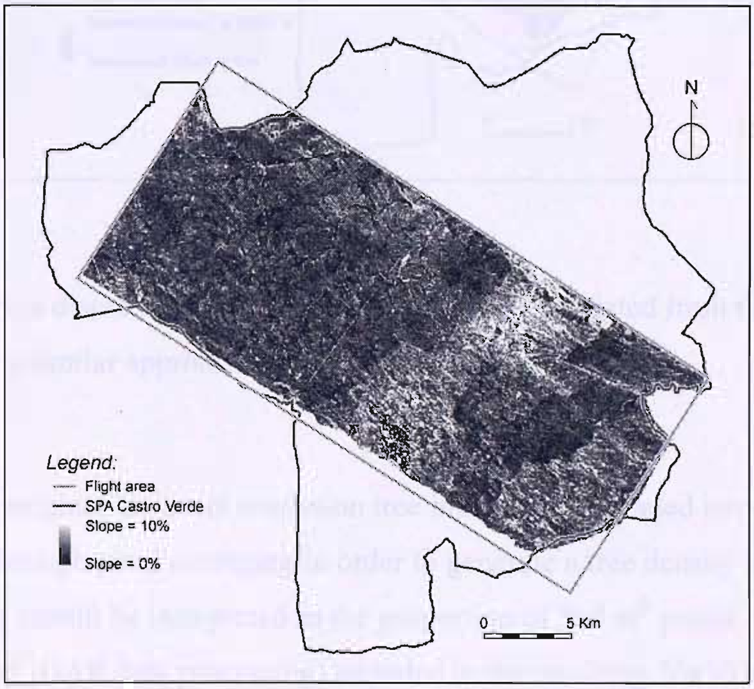


4.5. Environmental descriptors

The data layers resulting from the steps described above were then subsequently processed in order to generate useful environmental descriptors to be used in the SDMs. These were generated at a 30 m pixel resolution in order to match that of the Landsat-derived predictor variables (see Chapter 3).

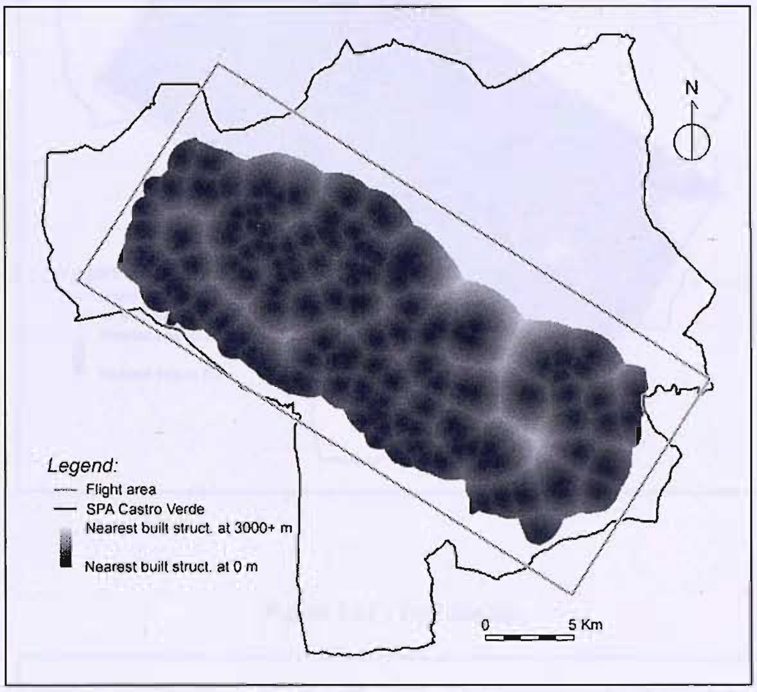
A Slope image was generated directly from the DTM. This was done by calculating, for each pixel, the tangent of the angle that has the maximum downhill slope in relation to the four neighbouring pixels (rook's case procedure), multiplied by 100 to produce a percentage gradient (Figure 4.8).

Figure 4.8 - Slope



A distance to the nearest built-up structure was calculated from the respective map of built-up structures. This reflects the Euclidian distance between each cell and the nearest of a set of target features. However, in order to guarantee a correct interpretation of this descriptor, the regions where the distance to the nearest feature was greater than the distance to the edge of the study area were eliminated (Figure 4.9). This variable is the one with the smallest extent, this way defining the maximum common area between all predictor variables at this scale of study.

Figure 4.9 - Distance to the nearest built-up structure


N
Legend:
Flight area
SPA Castro Verde
Nearest built struct. at 3000+ m
Nearest built struct. at 0 m
0
5 Km

Additionally, a distance to the nearest tree map was calculated from the map of trees, using a similar approach as described above (Figure 4.10).

Finally, the original 5m pixel resolution tree image was degraded into 30m pixel resolution through pixel averaging in order to generate a tree density map. This density map should be interpreted as the proportion of 5x5 m^2 pixels classified as trees (in the LiDAR data processing) included in the resulting 30x30 m^2 pixel (Figure 4.11).

Figure 4.10 - Distance to the nearest tree

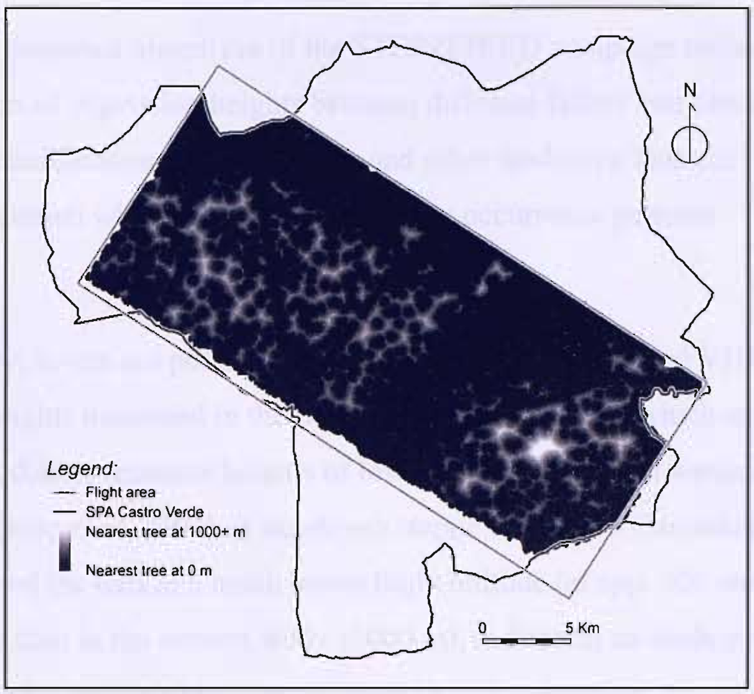
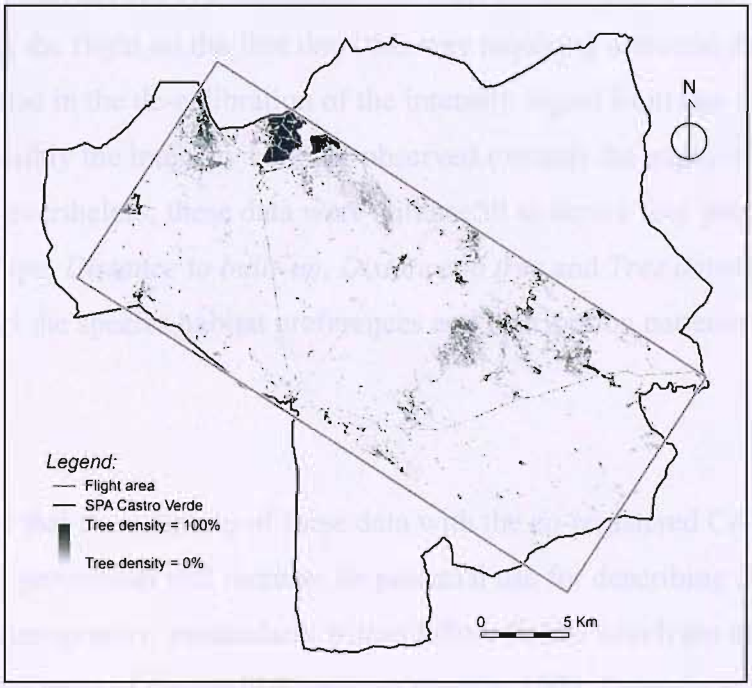


Figure 4.11 - Tree density



4.6. Final remarks

The initially proposed objectives of the STEPPEBIRD campaign included the discrimination of vegetation heights between different fallow and cereal fields, as well as the identification of shrubs, trees and other landscape features (such as built-up structures) which can influence species occurrence patterns.

Unfortunately, it was not possible to correlate the LiDAR-derived VHM with the vegetation heights measured in the field. Other recent studies which successfully used LiDAR data to measure heights of low vegetation, such as wetland vegetation (Genç *et al.* 2004) or sagebrush steppe vegetation (Streutker & Glenn 2006) collected the data at a much lower flight altitude (at app. 500 and 800 m, respectively) than in the present study (2000 m), indicating an inadequate mission planning for the objectives proposed in this campaign.

Additionally, the *ALTM 3033* sensor showed high instability, reflected in its failure during the flight on the first day (this way requiring a second day of flying), but also in the de-calibration of the intensity signal from one day to the next, and possibly the intensity banding observed towards the edges of each flight line. Nevertheless, these data were still useful to derive four predictor variables (*Slope*, *Distance to built-up*, *Distance to tree* and *Tree density*) to be used to model the species habitat preferences and distribution patterns within the study area.

It is expected that the coupling of these data with the co-registered CASI data (not yet fully processed) will increase its potential use for describing fine-scale landscape heterogeneity, particularly within fallow fields, which are of greatest importance for most of the studied species (Moreira 1999; Delgado & Moreira 2000).

5. Multi-scale analysis of steppe bird patterns of occurrence and habitat selection in Castro Verde

5.1. Introduction

As seen previously (Chapter 1), the birds of the steppe environments face a number of different threats relating to habitat degradation, such as agricultural intensification, land abandonment or afforestation (Tucker & Heath 1994; Burfield 2005; Santos & Suárez 2005). As a result, the vast majority of steppe bird species have unfavourable conservation status (BirdLife International 2004), which highlights their importance for biodiversity conservation. Nevertheless, conservation efforts require an understanding of species habitat preferences, as well as their patterns of distribution. Hence, many recent studies have focussed on the ecology of these birds, either at the community level (Moreira 1999; van Heezik & Seddon 1999; Delgado & Moreira 2000; Brotons *et al.* 2004a; Moreira *et al.* 2007; Traba *et al.* 2007; Oparin 2008) (see Appendix A.1) or at the species level (Martínez 1994; Stoate *et al.* 2000; Lane *et al.* 2001; Franco & Sutherland 2004; Pinto *et al.* 2005; Limiñana *et al.* 2006; Seoane *et al.* 2006).

None of these studies, however, has dealt with the issue of spatial scale and thus researched steppe bird habitat preferences or the observed occurrence patterns at different scales. Scale, though, is a fundamental issue in the spatial and environmental sciences, as all ecological processes possess an inherent scale at which they occur (Wiens 1989; Levin 1992; Whittaker *et al.* 2001; Blackburn & Gaston 2002). For this reason, a better understanding of species responses to the environment across different scales allows the development of better management measures, therefore having great importance for biodiversity conservation. The present work builds on this rationale and attempts to fill this research gap by investigating the habitat selection and distribution patterns of steppe birds in Castro Verde (a well-preserved pseudo-steppe region in South Portugal), at two different spatial scales. Additionally, a hierarchical multi-scale analysis is performed on some of the studied species (O'Neill *et al.* 1989; Pearson *et al.* 2004).

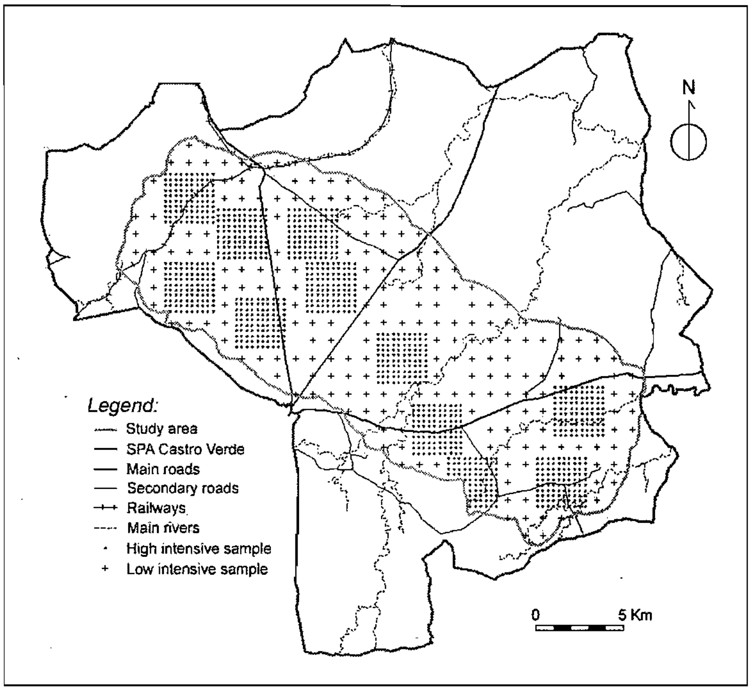
For this purpose, a SDM approach was used, which by quantifying the associations between the species and the environment, is suitable for the inference of habitat selection as well as for the prediction of the resulting occurrence patterns (Guisan & Zimmermann 2000; Suárez-Seoane *et al.* 2002a). Furthermore, SDMs can be applied at multiple spatial scales (Whittingham *et al.* 2005; Coreau & Martin 2007; Barbaro *et al.* 2008) and be used to support conservation planning (Ferrier 2002; Guisan & Thuiller 2005). Moreover, the landscape scale analysis here presented builds on a recent paper by Moreira *et al.* (2007) (see Appendix A.1) which provides the first results of an extensive survey of the steppe bird community within the pseudo-steppe areas of Castro Verde, carried out in parallel with the present work.

5.2. Methodology

- *Study area*

This study was carried out at two different spatial scales, with respectively different spatial extents. The regional study was conducted in the Baixo Alentejo region in southern Portugal, as described in previous chapters (see Chapter 1.2; Chapter 2.3; Figure 1.1). The landscape study is located within the previous region and inside the Castro Verde SPA. The spatial extent of the study at this scale is defined by the maximum common area between all descriptive variables (see Chapter 4.5; Figure 4.9; Figure 5.1).

Figure 5.1 - Landscape scale study area within the Castro Verde SPA and respective data sampling



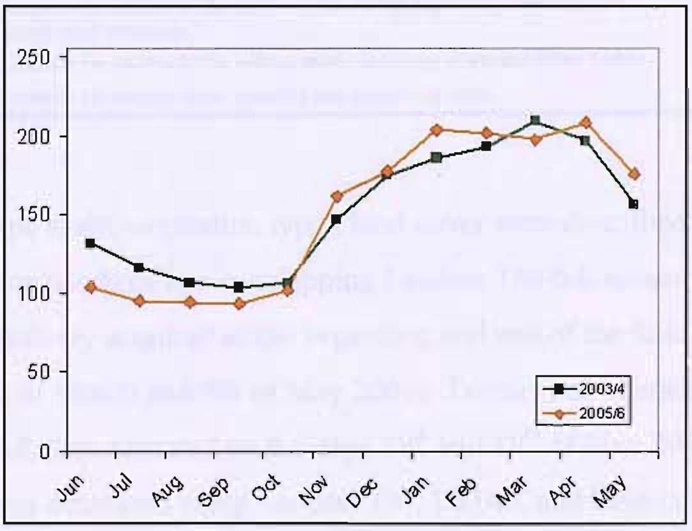
- *Species locational datasets*

The regional scale species dataset consisted of 557 occurrence (presence / absence) records, according to an intensive (geographically) stratified sampling scheme. These data were collected in the field during the spring of 2004 following the methodology described in Chapter 2.3. The landscape scale dataset included 1293 species occurrence records, according to combined systematic and random schemes, as follows. A portion of the dataset used (217 data records) follow a low-intensity sampling regular grid covering the full study area (see Appendix A1). The remaining of the data (1076 records) are located in a regular grid within 11 high-intensity sampling squares, which were randomly allocated over the region's steppe mosaic (see Chapter 3.2). All landscape data were collected during the spring of 2006 using a common field methodology (described in Chapter 3.2). In total, 16 species were considered, 11 of which are common to the studies at the two scales (Table 1.1). Both regional and landscape scale studies included 13 species each, and six species were used in the MS analysis.

- *Environmental descriptors*

Environmental descriptive data were used from several different sources at both study scales, but with a large emphasis on RS data. Following work by Osborne *et al.* (2001) and Suárez-Seoane *et al.* (2002a) it is assumed that steppe birds respond to factors relating to vegetation type / land cover, terrain and disturbance. At the regional scale, vegetation was described by two 12-month series of NDVI data from the SPOT VEGETATION sensor, for the periods of June 2003 to May 2004 and June 2005 to May 2006, respectively covering the year preceding the field campaigns at both study scales (Figure 5.2).

Figure 5.2 - SPOT VGT NDVI annual time-series for the periods 2003/4 and 2005/6



Terrain and disturbance were characterised as in Chapter 2.3. This chapter also describes the procedures for feature extraction of all regional scale data in order to derive the respective model predictor variables (Table 5.1). Additionally, an extra variable, quantifying the distance to the nearest river or water body (*Waterdist*), was included as a disturbance factor. This was derived from a hydrographical map provided by the Agência Portuguesa do Ambiente / Atlas do

Ambiente Digital (http://www2.apambiente.pt/atlas/est/index.jsp), using the same methodology as used for extracting *Roaddist*.

Table 5.1 - Predictor variables used in the regional scale models

Variable	Description	Data source
Vegetation		
Summer	Vegetation senescence during the Summer months: *Summer = NDVI (Jun) – NDVI (Sep)*	SPOT VGT
Winter	Vegetation growth during the Autumn and Winter months: *Winter = NDVI (Mar) – NDVI (Sep)*	SPOT VGT
Spring	Vegetation senescence during the Spring months: *Spring = NDVI (Jan) – NDVI (Apr)*	SPOT VGT
Dry	Mean NDVI during the dry months: *Dry = Average [NDVI (Jun : Oct)]*	SPOT VGT
Wet	Mean NDVI during the wet months: *Wet = Average [NDVI (Jan : Apr)]*	SPOT VGT
Dec	NDVI value for the month of December: *NDVI (Dec)*	SPOT VGT
May	NDVI value for the month of May: *NDVI (May)*	SPOT VGT
Terrain		
Alt	Mean altitude in metres within a 5 x 5 array of 200 x 200 m pixels	DTM
Topov10	Variation in altitude in a 5 x 5 array of 200 x 200 m pixels, where altitude is re-classed to a 10 m vertical resolution. *Topov10 = (n-1)/(p-1), where n = number of different altitude classes in the array, p = number of pixels in the array, i.e. 25*	DTM
Disturbance		
Urbandist	Distance (in metres) to the nearest pixel containing towns, settlements or constructed structures	CORINE LC
Waterdist	Distance (in metres) to the nearest pixel containing rivers and water bodies	River map
Roaddist	Distance (in metres) to the nearest pixel containing roads	Road map

At the landscape scale, vegetation type / land cover were described by data from the six reflective bands of two overlapping Landsat TM full scenes (path/row: 203/34), respectively acquired at the beginning and end of the field data collection (6th of March and 9th of May 2006). Terrain was characterized by airborne LiDAR data acquired on the days 18th and 19th of May 2006. Disturbance was described using Landsat TM, LiDAR and field-collected GPS data. At this scale, all predictor variables were originally generated at a pixel resolution of 30 x 30 m, and the model training dataset was generated by synthesizing these data within the field sampled circular-plots (see Chapter 3.2; Table 5.2). All Landsat TM data processing for feature extraction is described in Chapter 3.2, being the resulting vegetation type / land cover predictor variables resulting from the SVM soft classifications. The use of soft classification outputs should allow for an ecological interpretation of the species habitat preferences while keeping the numerical heterogeneity of the spectral data (see Chapter 3.4). The disturbance variable *D2water*, however, used the map resulting from the

SVM hard classification of the March imagery. The feature extraction of the LiDAR-derived predictors is described in Chapters 4.4 and 4.5. Additionally to these, the variable *Terrainvar* was calculated as the standard deviation of *Slope*, within the respective circular-plot. The *Dist2road* variable was derived from a GPS data coverage of the paved roads in the study area. Finally, the prediction dataset was generated by applying a moving-window kernel filter (describing the circular-plot) to the predictor layers for the full study area. This procedure generated a new set of predictor variable layers with the same 30m x 30m pixel resolution, but containing information relative to the 125 m radius circular-plot around each pixel centre, i.e. the grain of analysis of the trained models.

Table 5.2 - Predictor variables used in the landscape scale models

Variable	Description	**Data source**
Vegetation type / Land Cover		
BS_Mar	Bare soil or nor vegetation in March, including ploughed fields and dirt tracks and paved / built-up areas	Landsat TM
C_Mar	Cereal and forage crop fields in March, possibly including some grass-dominated fallows	Landsat TM
BS_May	Bare soil or nor vegetation in May, including ploughed fields and dirt tracks and paved/built-up areas	Landsat TM
Fallow	Fallow fields, except areas included in "*DV_May*"	Landsat TM
C_May	Cereal and forage crop fields in May, possibly including some grass-dominated fallows, except areas included in "*DV_May*"	Landsat TM
DV_May	Phenological gradient from green to dry (senescent) vegetation, mostly relevant for herbaceous vegetation (Gramineae) such as cereal / forage fields or grass-dominated fallows	Landsat TM
WdShr	All areas of woodland and shrubs	Landsat TM
Terrain		
Slope	Surface slope (in percentage)	LiDAR
Terrainvar	Terrain variability: standard deviation of "*Slope*"	LiDAR
Disturbance		
D2water	Distance (in metres) to the nearest pixel containing water bodies	Landsat TM
D2road	Distance (in metres) to the nearest pixel containing paved roads	GPS
D2built	Distance (in metres) to the nearest pixel containing built-up structures	LiDAR
D2tree	Distance (in metres) to the nearest pixel containing trees	LiDAR
Treedens	Tree density: proportion of 5 m pixels classified as "tree" within each 30m pixel	LiDAR

- *Model building*

All models were built according to a common methodological framework, at both scales of study. The species occurrence data were fitted to the environmental descriptors using MARS models (Friedman 1991), for the purposes of prediction of the occurrence patterns and inference of habitat preferences. These models were implemented in R statistical software (R Development Core Team 2008) using code from the *mda* package (Hastie & Tibshirani 1996) further modified to allow for binary data (using a logit link function) and n-fold model cross-validation (Elith & Leathwick 2007).

A detailed description of the functioning of MARS models is included in Chapter 3.2. They are considered to be fast and high performing (De Veaux *et al.* 1993; Elith *et al.* 2006), although their vulnerability to high collinearity (or concurvity) in the predictor data can potentially result in performance loss as well as present problems in model interpretability (Friedman 1991; De Veaux & Ungar 1994; Morlini 2006). Also, the effects of data collinearity on the models, by relating to the particular associations between the response variable and the respective inter-correlated predictors, are expected to be case-specific (Snee & Marquardt 1984). Moreover, data multi-collinearity, by affecting the model variable selection procedure, is difficult to guard against in one-model approaches (MacNally 2000). For this reason, and in order to account for these varying effects, three models were run for each species, using predictor subsets defined according to a varying maximum (Spearman) rank correlation between predictor variables, set respectively at 0.7, 0.6 and 0.5. The selection of the respective sets of predictor variables was species-specific and done through a variable reduction approach as follows. In cases where two predictor variables were correlated above the set threshold, the one that best fitted (in terms of deviance explained) the respective response variable, using univariate MARS with a minimal backfitting penalization (penalty = 0), was selected. Similarly, in cases where more than two predictor variables were highly inter-correlated, all possible sets of variable combinations (according to the criterion) were examined and the set with the highest fit to the response was selected. Subsequently, the data were fitted using single-response MARS models (for each species). The selection of the best model parameters (of maximum interaction order “mars.degree” and backfitting penalization “penalty”) was done through a grid search procedure (by searching all possible parameter combinations, in a grid), this way optimising the parameters according to model performance, specific to each dataset. In this search, the maximal interaction order parameter was allowed to assume values of 0 (additive model) and 1 (allowing only first order interactions), and the backfitting penalization parameter could vary from 0 (minimal penalization) to 4 (maximal penalization), in discrete steps. Model performance was assessed through a 10-fold cross-validation procedure, while controlling for prevalence in the data resampling. On each case, five replications were performed to account for the variability inherent in the cross-validation process (Breiman 1996). The averaged (out of the five replicates) ROC AUC scores (Hanley & McNeil 1982) were used as a model performance measure (*AUCcv*). Additionally, the variables' drop contributions of the selected models (on each species-predictor dataset) were calculated in terms of percentage loss in model deviance explained when excluding each variable. Also, the model univariate partial regression plots were extracted. These represent the species response curves to each variable (or response shapes in the case of interacting variables) in the model. Additive and interacting effects between variables can result in multivariate responses different to those expressed in (single variable) partial plots, and it is therefore important to observe the response curves / shapes for the final model to ensure that they remain ecologically reasonable (Wintle *et al.* 2005).

The overall best scoring model for each species (at each scale) was used for predicting its probability of occurrence within the (respective) study area, conditional on a minimal model performance (*AUCcv*) value of 0.7, indicative of "good" model performance (*sensu* Hosmer & Lemeshow 2000). For visualisation purposes, the landscape scale predictions (originally at a 30 x 30 m pixel resolution, but containing neighbouring information relative to an area equivalent to the data sampling circular-plot) were degraded to a 100 x 100 m pixel resolution, by pixel averaging. In order to assess the species preference (at the regional level) for the Castro Verde SPA, the averaged (regional) model prediction values were compared between pixels within this region and the whole Baixo Alentejo.

For the purpose of inference of the species-environment associations, however, the use of the single best model is not appropriate due to the of data collinearity on the models' variable selection procedure. MacNally (2000) suggests that single-model approaches are difficult to guard against multi-collinearity, making them ineffective for identifying those variables most likely to influence variation in the dependent (response) variable. Thus, the three selected models (with the different predictor subsets) for each species were considered. The average model variable contributions were calculated for each case and the plots of the most contributing variables (on average) were inspected. From these plots, the typical (consistent among all three selected models) responses were identified and interpreted. This procedure aimed at identifying the predictor variables that mostly influenced the models, across different levels of data collinearity. Species responses were interpreted through the use of model partial plots, which represent the relationship between each variable and the probability of occurrence of the respective species, in the multivariate model context, independent of the other variables included in the model (Wintle *et al.* 2005).

The MS models were built in a hierarchical manner, using a top-down approach, by incorporating the predictions from the regional scale models as descriptive variables to be used as predictors at the landscape scale (O'Neill *et al.* 1989; Turner *et al.* 1989; Pearson *et al.* 2004). In this way it was possible to assess if the species' regional scale occurrence patterns can aid the explanation of those observed at the landscape level. The regional scale models, however, were trained on datasets relative to the spring of 2004, different to that of the landscape scale data (spring of 2006). Therefore, these models were applied to regional scale datasets collected at both periods, thus, generating predictions of the occurrence patterns of steppe birds in the region for each year. These were then used separately in different MS models, respectively M2004 and M2006. Even though the main annual phenological patterns of the vegetation observed in the study region were similar in both years (Figure 5.2), the correlations between model predictors between years was always relatively low, with maximum value of 0.779, for variable *Dry* (Table 5.3), although always highly significant ($p < 0.0001$). Moreover, a GLM Analysis of Variance (ANOVA) of these datasets showed highly significant differences in the distribution of the data values between different years on all variables except for *Winter* (Table 5.3). Nevertheless, it was assumed that the species-environment associations fit on the 2004 training dataset are transferable to 2006 and to the full regional study area. The use of the regional model predictions at the landscape scale, though, assumes only transferability within the Castro Verde study area.

Table 5.3 - Comparison of vegetation descriptors for the Baixo Alentejo between both years: Pearson r correlation coefficient; and p value of significance of the GLM ANOVA

Variable	Pearson r	ANOVA p
Summer	0.325	0.0001
Winter	0.626	0.097
Spring	0.498	0.0001
Dry	0.779	0.0001
Wet	0.450	0.0001
Dec	0.311	0.0001
May	0.545	0.0001

Hence, multi-scale MARS models were fitted only for those species with data collected at both spatial scales and with the respective regional model performance (*AUCcv*) above 0.7. The respective model parameters were optimized for performance by using the same grid search approach as in the single-scale models. In addition, the variable drop contributions were calculated for the selected models, and the respective partial plots extracted. Finally, the Kendall's τ (*tau*) rank correlation coefficient was calculated for the variable (drop) contribution values between the models at the landscape scale and the respective MS models. This measure is used for assessing model consistency, as an indicator of the effect of the introduction of the regional prediction variable in model structure.

5.3. *Results*

- *Species distribution models*

At each study scale, three models were selected for each species, one for each predictor dataset, with maximum rank correlations between variables set respectively at 0.7, 0.6 and 0.5. The performance scores (*AUCcv*) of the selected regional scale models varied between 0.590 and 0.807 (mean = 0.725; SE = 0.018) when using a correlation threshold of 0.7 (R70), between 0.580 and 0.836 (mean = 0.728; SE = 0.019) with a threshold of 0.6 (R60) and between 0.579 and 0.813 (mean = 0.725; SE = 0.019) by setting the maximum correlation between variables to 0.5 (R50) – see Table 5.4.

Table 5.4 - Species frequency of occurrence and selected regional scale models, when using rank correlation thresholds of 0.70 (R70), 0.60 (R60) and 0.50 (R50) between input variables (best performing models in bold); model parameters are expressed in the form « "mars.degree". "penalty" »

Species	**Freq. of occurr.**	**R70**		**R60**		**R50**	
		Model	***AUCcv***	***Model***	***AUCcv***	***Model***	***AUCcv***
Cirpyg	0.188	1.4	0.771	1.3	0.771	**1.1**	**0.780**
Aleruf	0.519	**2.0**	**0.682**	1.3	0.673	1.0	0.671
Cotcot	0.461	2.3	0.754	1.3	0.746	**2.2**	**0.756**
Tettet	0.285	1.2	0.784	1.0	0.783	**1.0**	**0.793**
Otitar	0.058	**1.1**	**0.807**	1.2	0.792	2.1	0.787
Buroed	0.032	1.0	0.695	**1.0**	**0.699**	2.1	0.675
Pteori	0.022	1.2	0.799	**1.4**	**0.836**	1.4	0.813
Galsp	0.232	1.3	0.675	**1.3**	**0.682**	2.0	0.652
Melcal	0.132	1.3	0.770	**2.1**	**0.789**	1.0	0.780
Oenhis	0.055	1.4	0.658	**1.4**	**0.673**	1.4	0.671
Saxtor	0.440	**2.0**	**0.590**	2.3	0.580	2.1	0.579
Cisjun	0.637	**1.1**	**0.738**	1.1	0.733	1.1	0.725
Milcal	0.903	1.3	0.699	1.0	0.706	**1.4**	**0.739**

The best performing models were achieved for four of the studied species (*Aleruf*, *Otitar*, *Saxtor* and *Cisjun*), for five species (*Buroed*, *Cotcot*, *Pteori*, *Galsp* and *Oenhis*) and for four species (*Cirpyg*, *Cotcot*, *Tettet* and *Milcal*), respectively with datasets R70, R60 and R50. The performance of all best models (one per species) ranged between 0.590 and 0.836 (mean = 0.736; SE = 0.019). The maximum model performance difference between datasets of the same species varied between 0.009 and 0.040 (mean = 0.019; SE = 0.003). The selected models included variable interactions (mars.degree = 2) for three, two and five species, respectively on the referred datasets. However, only for one species (*Saxtor*) were variable interactions consistently used on all three selected models, and on two of the selected models for one species (*Cotcot*).

In terms of model structure, the variable that overall (on average, between the three subsets, and across species) most contributed in the models was *Dry*. In fact, this was the variable with the highest drop contribution in the models for five of the studied species (*Cirpyg*, *Tettet*, *Otitar*, *Pteori* and *Melcal*) and was highly contributing in the models of two other species (*Buroed* and *Galsp*) – see Table 5.5. Variables *May* and *Wet* also had overall large contribution in the species models, being the most contributing for two (*Aleruf* and *Buroed*) and for three species (*Cotcot*, *Galsp* and *Saxtor*), respectively. Additionally, *Spring*, *Dec* and *Topovl0* were the variables with the highest drop contribution for one species each (respectively *Oenhis*, *Cisjun* and *Milcal*).

Table 5.5 - Mean relative drop contribution of the regional scale models' predictor variables (in proportion of the deviance explained by the model); the most contributing variables for each species are in bold

	Summer	***Winter***	***Spring***	***Dry***	***Wet***	***Dec***	***May***	***Alt***	***Topovl0***	***Roaddist***	***Waterdist***	***Urbandist***
Cirpyg	0	0	0	**0.435**	0	0	0.144	0.108	0	0	0	0
Aleruf	0.144	0.138	0	0.040	0.052	0	**0.456**	0.101	0	**0.204**	0	0.051
Cotcot	0.010	0	0.169	0	**0.208**	0.128	0.130	0.055	0.169	0	0	0
Tettet	0.059	0	0	**0.407**	0.150	0	0.050	0	0.045	0.009	0.031	0
Otitar	**0.216**	0	0.007	**0.490**	0.027	0.065	0.016	0	0	0.004	0	0
Buroed	**0.243**	0	0	**0.202**	0.039	0	**0.311**	0.054	0	0	0	0.105
Pteori	0	0	0.004	**0.757**	0.036	0.151	0	0	0	0	0.043	0
Galsp	0.090	0.199	0	**0.351**	**0.356**	**0.306**	0	0	0.009	0	0.074	0
Melcal	0.064	0	0	**0.561**	0.024	0.035	0	0.123	0.019	0	0.017	0.032
Oenhis	0	0	**0.667**	0	0	0	**0.333**	0	0	0	0	0
Saxtor	0.143	**0.181**	0	0.100	**0.318**	0.085	0.099	0.030	0.123	0.023	0.165	0.016
Cisjun	0.068	0.114	0	0	0.148	**0.290**	0.150	0.028	0.010	0	0	**0.203**
Milcal	0.049	0	0	0.025	**0.225**	0.061	0	0.018	**0.391**	0	0	0
MEAN	0.084	0.049	0.065	**0.259**	**0.122**	0.086	**0.130**	0.040	0.059	0.018	0.025	0.031

At the landscape scale, datasets obtained by setting variable correlation thresholds at 0.7 (L70), 0.6 (L60) and 0.5 (L50) generated models with performance scores ranging respectively from 0.571 to 0.836 (mean = 0.711; SE = 0.023), from 0.576 to 0.840 (mean = 0.708; SE = 0.022) and from 0.586 to 0.833 (mean = 0.715; SE = 0.022) – see Table 5.6. These datasets resulted in the best performing models for four (*Cirpyg*, *Melcal*, *Cisjun* and *Milcal*), one (*Galsp*) and eight species (*Aleruf*, *Tettet*, *Otitar*, *Buroed*, *Calbra*, *Antcam*, *Oenhis* and *Saxtor*), respectively. The performance of the best models for all species ranged between 0.586 and 0.840 (mean = 0.718; SE = 0.022). The maximum performance difference observed between datasets of each species varied between 0.002 and 0.033 (mean = 0.013; SE = 0.003). The selected models included interactions between variables for two species on each of the datasets. Consistency in the use of variable interactions (on all datasets) was, however, only observed on one species (*Tettet*).

Table 5.6 - Species frequency of occurrence and selected landscape scale models, when using rank correlation thresholds of 0.70 (L70), 0.60 (L60) and 0.50 (L50) between input variables (best performing models in bold); model parameters are expressed in the form « "mars.degree". "penalty" »

Species	**Freq. of occurr.**	**L70**		**L60**		**L50**	
		Model	***AUCcv***	***Model***	***AUCcv***	***Model***	***AUCcv***
Cirpyg	0.062	**1.3**	**0.630**	1.4	0.621	1.0	0.624
Aleruf	0.077	1.3	0.674	1.2	0.676	**1.2**	**0.683**
Tettet	0.160	2.2	0.682	2.0	0.681	**2.2**	**0.683**
Otitar	0.052	2.1	0.571	1.0	0.576	**1.0**	**0.586**
Buroed	0.032	1.3	0.595	1.4	0.585	**1.3**	**0.599**
Galsp	0.232	1.1	0.836	**1.0**	**0.840**	1.0	0.833
Melcal	0.283	**1.0**	**0.803**	1.1	0.799	1.0	0.800
Calbra	0.102	1.4	0.758	2.2	0.757	**1.2**	**0.762**
Antcam	0.028	1.4	0.737	1.4	0.729	**1.3**	**0.749**
Oenhis	0.033	1.0	0.686	1.1	0.708	**1.0**	**0.719**
Saxtor	0.110	1.2	0.731	1.1	0.735	**1.1**	**0.738**
Cisjun	0.114	**1.4**	**0.806**	1.3	0.780	1.4	0.800
Milcal	0.778	**1.4**	**0.735**	1.4	0.720	2.2	0.716

The variables *C_Mar* and *BS_May* were the two overall most contributing in the landscape models, being the ones with the highest drop contribution scores (averaged between the three datasets) respectively for five (*Cirpyg*, *Tettet*, *Melcal*, *Calbra* and *Cisjun*) and three of the studied species (*Otitar*, *Buroed* and *Oenhis*) – see Table 5.7. Furthermore, *Dist2tree* was also the most contributing variable for three species (*Galsp*, *Melcal* and *Saxtor*) even though with a much lower overall mean contributing score as the previous. Also, both *BS_Mar* and *Slope* were the variables with the highest drop contribution for one species each (respectively *Antcam* and *Aleruf*).

Table 5.7 - Mean relative drop contribution of the landscape scale models' predictors (in proportion of the deviance explained by the model); the most contributing variables for each species are in bold

	BS_Mar	***C_Mar***	***BS_May***	***Fallow***	***C_May***	***DV_May***	***WdShr***	***Slope***	***Terrainvar***	***D2water***	***D2road***	***D2built***	***D2tree***	***Treedens***
Cirpyg	0	**0.487**	0	0.047	0	0.047	0	0	0	0	0	0	0.097	0
Aleruf	0.099	**0.161**	0.022	0	0	0	**0.179**	**0.241**	0	0.101	0.019	0	0	0
Tettet	0.024	**0.553**	0	0.051	0	0	**0.137**	0	0	**0.238**	0	0.051	0	**0.158**
Otitar	0	0.015	**0.528**	0.111	0	0.026	0.045	0.024	0	0.093	0	**0.163**	0	0
Buroed	0	0	**0.755**	0	0	0	**0.096**	0	0	0	0	0	0	0
Galsp	0.022	0.045	0.018	0	0.033	0.006	0.010	0	0.053	0.072	0.057	**0.080**	**0.096**	0
Melcal	0	**0.199**	0.025	0.024	0.014	0.022	0.057	0	0.044	0.032	0.028	0.060	**0.164**	0
Calbra	**0.134**	**0.217**	**0.167**	0	0.012	0	0	0	0	0	0	0.006	**0.134**	0
Antcam	**0.572**	0	0	0	0	0	0	0	0	0.150	0	0	0	0
Oenhis	0	0.064	**0.381**	0.146	0	0	0.134	0.113	0	0	**0.160**	0.071	0	0
Saxtor	0.069	0	0	0	0.055	0.080	**0.177**	0	0.067	0	0.099	0.062	**0.190**	0
Cisjun	0.009	**0.319**	0.033	**0.242**	0.037	0	0.035	0	0.007	0	0.048	0	0	0
Milcal	**0.296**	0.017	0.062	**0.233**	**0.199**	0.123	0.054	0	0	0	0	0.039	0	0
MEAN	**0.094**	**0.160**	**0.153**	0.066	0.027	0.023	**0.071**	0.029	0.013	**0.053**	0.032	0.041	**0.052**	0.012

Of the eleven species in common between the two scales, the regional scale models best fitted the occurrence data of four of them (*Cirpyg*, *Tettet*, *Otitar* and *Buroed*), the landscape models those of five species (*Galsp*, *Melcal*, *Oenhis*, *Saxtor* and *Cisjun*), and on two species the regional and landscape scale models achieve similar performance values, i.e. with a performance difference smaller than 0.010 (Table 5.8).

Table 5.8 - Performance of the best selected model for each species at each spatial scale (in the cases where the performance difference between the two scales is greater than 0.010, the highest value is in bold)

Species	Regional	Landscape
Cirpyg	**0.780**	0.630
Aleruf	0.682	0.683
Cotcot	0.756	-
Tettet	**0.793**	0.683
Otitar	**0.807**	0.586
Buroed	**0.699**	0.599
Pteori	0.836	-
Galsp	0.682	**0.840**
Melcal	0.789	**0.803**
Calbra	-	0.762
Antcam	-	0.749
Oenhis	0.673	**0.719**
Saxtor	0.590	**0.738**
Cisjun	0.738	**0.806**
Milcal	0.739	0.735
MEAN	0.736	0.718

In the MS analysis, the selected models achieved performances ranging between 0.619 and 0.813 (mean = 0.717; SE = 0.023) and between 0.632 and 0.804 (mean = 0.723; SE = 0.020), respectively when using regional model predictions for 2004 (M2004) and 2006 (M2006) – see Table 5.9. The change in model performance of these models, when compared with the respective landscape scale models ranged between -0.010 and 0.053 (mean = 0.009; SE = 0.006) and between -0.004 and 0.080 (mean = 0.015; SE = 0.009), respectively for models M2004 and M2006. The regional predictions for 2004 were included as predictor variables in the selected models of four species (*Cirpyg*, *Otitar*, *Melcal* and *Milcal*), and those for 2006 in the models of three species (*Otitar*, *Melcal* and *Milcal*). Also, the performance change (compared with the respective landscape models) of those models which selected the regional predictions as predictor variables ranged respectively between -0.010 and 0.053 (mean = 0.014; SE = 0.008) and between -0.002 and 0.080 (mean = 0.032; SE = 0.012). The selected models included interactions between variables for one species (*Tettet*) and for two species (*Tettet* and *Milcal*), respectively on M2004 and M2006.

Table 5.9 - Selected multi-scale models, by using regional predictions for 2004 (M2004) and 2006 (M2006), respective model performance (AUCcv) and performance change (ΔAUCcv) when comparing with the respective landscape scale model; in bold, the models which selected the regional predictions; model parameters are expressed in the form « "mars.degree". "penalty" »

Species	**M2004**			**M2006**		
	Model	***AUCcv***	***ΔAUCcv***	***Model***	***AUCcv***	***ΔAUCcv***
Cirpyg	**1.1**	**0.619**	- 0.010	1.2	0.632	+ 0.002
Tettet	2.3	0.682	- 0.001	2.2	0.681	- 0.002
Otitar	**1.0**	**0.639**	+ 0.053	**1.0**	**0.666**	+ 0.080
Melcal	**1.1**	**0.813**	+ 0.007	**1.0**	**0.804**	- 0.002
Cisjun	1.4	0.803	- 0.003	1.4	0.803	- 0.004
Milcal	**1.0**	**0.743**	+ 0.008	**2.4**	**0.753**	+ 0.018

The regional predictions for 2004 were selected as predictor variables in the MS models on four of the six species, and those for 2006 on three species (Table 5.10). Moreover, the change in model performance, when compared with the respective landscape models, was found to be highly correlated with the drop contributions of the regional predictions in both the M2004 ($r = 0.954$; $n= 6$; $p < 0.01$) and M2006 models ($r = 0.992$; $n = 6$; $p < 0.0001$). In terms of model structure consistency between the landscape and MS models it was found that, similarly on M2004 and M2006, it was observed that the greater the regional predictions' contribution in the models the more different the respective model structures (smaller the Kendall τ), for three of the species (*Cirpyg*, *Tettet* and *Otitar*). On the other hand, the same pattern was not found for the remaining three species, where for *Melcal* and *Milcal* a relatively small contribution of the regional predictions coincide with a low τ score. Moreover, in the case of *Cisjun*, even though the regional model predictions were not selected in the final model, their presence in the predictor dataset was sufficient for the new models being built with a very different variable contribution structure.

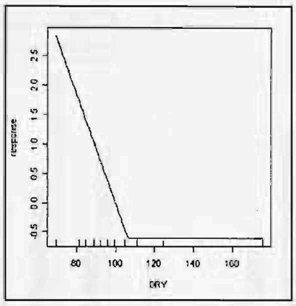
Table 5.10 - Drop contribution of the regional predictions in the multi-scale models (in proportion of the respective model deviance explained)

	M2004		M2006	
	Contrib.	***Kendall τ***	***Contrib.***	***Kendall τ***
Cirpyg	0.053	0.770	0	0.864
Tettet	0	0.812	0	0.910
Otitar	**0.314**	0.737	**0.462**	0.371
Melcal	0.075	0.613	0.024	0.409
Cisjun	0	0.634	0	0.497
Milcal	0.063	0.585	0.158	0.535

- *Habitat selection and predicted occurrence patterns: species accounts*

At the regional scale, Montagu's Harrier (*Cirpyg*) mostly responded to the variable *Dry*, by selecting areas with respective low values (Table 5.5; Figure 5.3). These areas should reflect the generally drier or less vegetated areas during the dry and warm months, which should roughly correspond to the low-intensity agricultural areas, such as fallow grasslands or dry cereal crops (as opposed to e.g. irrigated crops, shrublands or forests). Other lower contributing variables in the selected models were *May* and *Alt.* No disturbance variables were used by these species' models.

Figure 5.3 - Regional scale model partial plots for Montagu's Harrier (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



This species is predicted to occur mainly in the southern part of the Baixo Alentejo region, though avoiding the Serra de Caldeirão hills and the Guadiana valley (see Figure 5.4). Within the Castro Verde SPA, the species avoids the non-steppe areas (see Figure 3.1) and the valleys containing the two main rivers (Figure 1.2). By comparing the predicted probability values (out of the regional models) for the Baixo Alentejo region and for the Castro Verde SPA (see Figure 5.5), it is possible to observe a clear preference of this species for the latter – with a mean predicted probability value in Castro Verde of more than double (237.4 %) of that for the whole region.

Figure 5.4 - Predicted probabilities of occurrence of Montagu's Harrier in the Baixo Alentejo

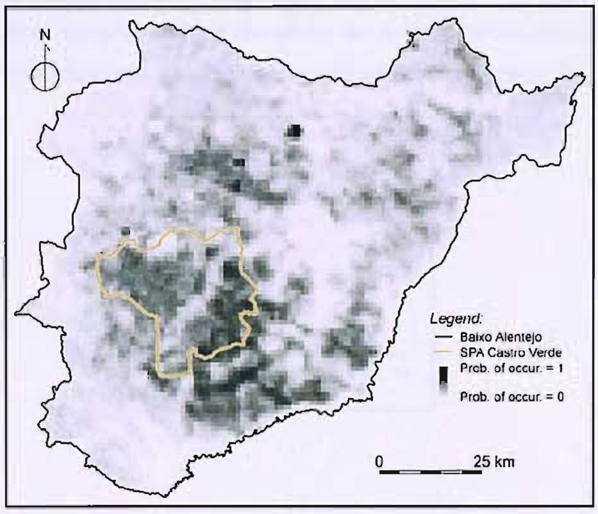
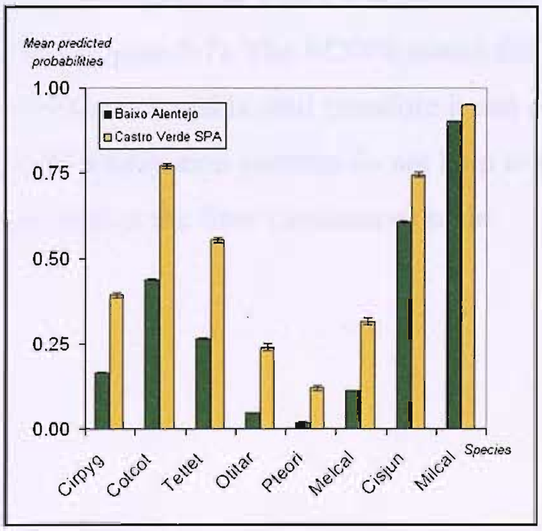
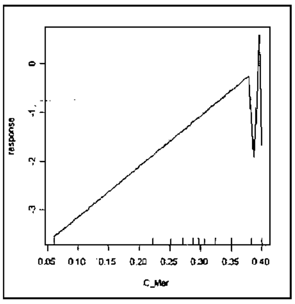


Figure 5.5 - Mean values of predicted probability (at the regional scale) in the whole Baixo Alentejo region and within the Castro Verde SPA



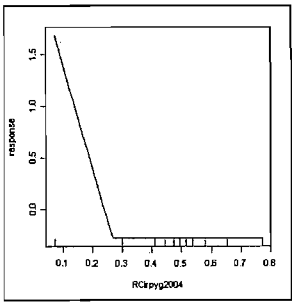
The landscape scale models for this species never achieved "good" performance values (*sensu* (Hosmer & Lemeshow 2000), and thus the occurrence patterns of Montagu's Harrier for the Castro Verde SPA at this scale were not predicted. Nevertheless, the species data were fitted mostly by the variable *C_Mar*, suggesting its relevance for the species (Table 5.7). This variable describes the fields cultivated with cereal early in the breeding season, which seems to be favourable for the occurrence of these birds (Figure 5.6). Minor contributing variables on the selected models were *D2tree*, *Fallow* and *DV_May*.

Figure 5.6 - Landscape scale model partial plots for Montagu's Harrier (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



The M2004 multi-scale model for this species selected the regional predictions as a predictor variable, even though this resulted in a model performance loss. Nevertheless, by inspecting the species response to this variable, it was observed that the species (in 2006) selected areas with a low probability of occurrence at the regional scale in 2004 (Figure 5.7). The M2006 model did not select the regional prediction as predictor variable, and therefore it can only be assumed that the predicted regional distribution patterns do not help to explain the observed occurrence patterns at the finer (landscape) scale.

Figure 5.7 - MS model partial plots for Montagu's Harrier: M2004 (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



The selected Red-legged Partridge (*Aleruf*) models never achieved "good" performance values, at either of the two scales, hence no predictions of occurrence patterns were generated for this species. The regional scale models (Table 5.4), however, had greater contributions from the variables *May* and *Roaddist* (Table 5.5). While the fitted responses to *May* were difficult to interpret (including interactions with *Summer* and *Wet* on model R70) the response to *Roaddist* showed an avoidance of areas closer than 2000 m from the nearest paved road (Figure 5.8). Additionally, other variables used by the selected models, though with lower (drop) contributions were *Summer*, *Winter*, *Alt*, *Wet*, *Urbandist* and *Dry*.

The landscape scale models (Table 5.6), on the other hand, had greater contributions from variables *Slope*, *WdShr* and C_*Mar* (Table 5.7). The examination of the model partial plots indicated that this species selected areas with higher slope and with woodland and shrubs, while avoiding areas without cereal fields (Figure 5.9). Minor contributing variables in the selected models included *D2water*; *BS_Mar*, *BS_May* and *D2road.*

Figure 5.8 - Regional scale model partial plots for Red-legged Partridge (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)

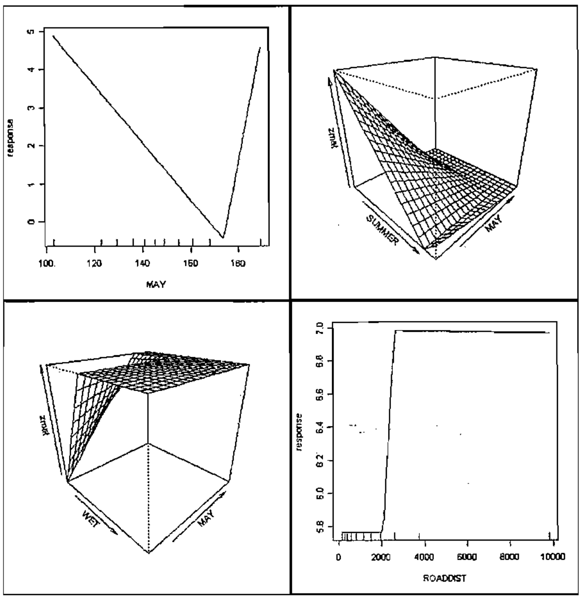
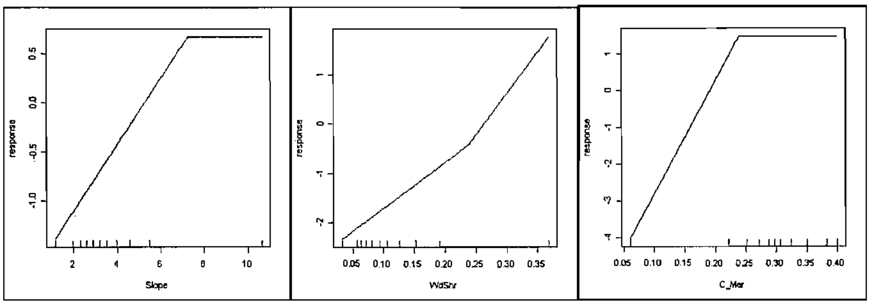


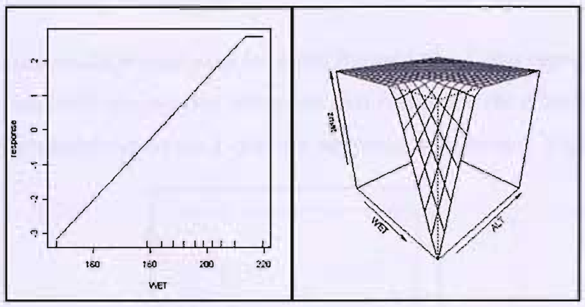
Figure 5.9 - Landscape scale model partial plots for Red-legged Partridge (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



Quail (*Cotcot*) occurrence data was only collected at the regional scale and thus no landscape scale models were built. Variable *Wet* was the most contributing variable in the selected models for this species (Table 5.4; Table 5.5), with greater probabilities of occurrence in areas with high NDVI values during the months of January to April (Figure 5.10), which should correspond to areas of greater winter cereal production. Two of the selected models (R70 and R50) included interactions between this variable and *Alt*, by avoiding areas with high values of *Wet* and low altitude. Less contributing variables in the models were

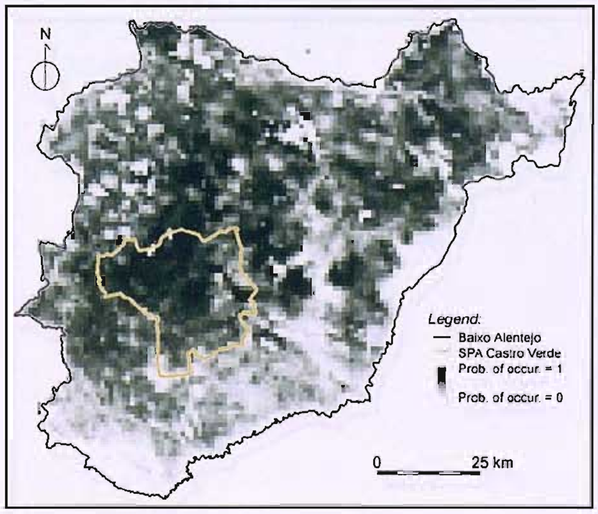
Spring, *Topov10*, *May*, *Dec*, *Alt* and *Summer*. These species data, therefore, was never fitted with any disturbance variables.

Figure 5.10 - Regional scale model partial plots for Quail (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



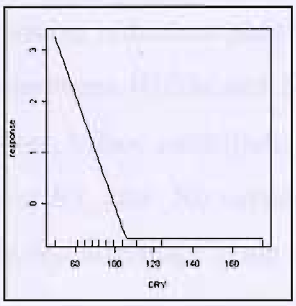
The predicted occurrence patterns of Quail show a generalised distribution of the species throughout the Baixo Alentejo region, possibly reflecting the distribution of the winter cereal areas (Figure 5.11). Within the Castro Verde SPA, the species is predicted to occur over most of the area, though avoiding some regions of less favourable habitat such as the holm oak 'montado' areas at the S and NE of the area, as well as some existing afforestations. The mean regional prediction value for the Castro Verde SPA was 153.5 % of that for the Baixo Alentejo region (see Figure 5.5).

Figure 5.11 - Predicted probabilities of occurrence of Quail in the Baixo Alentejo



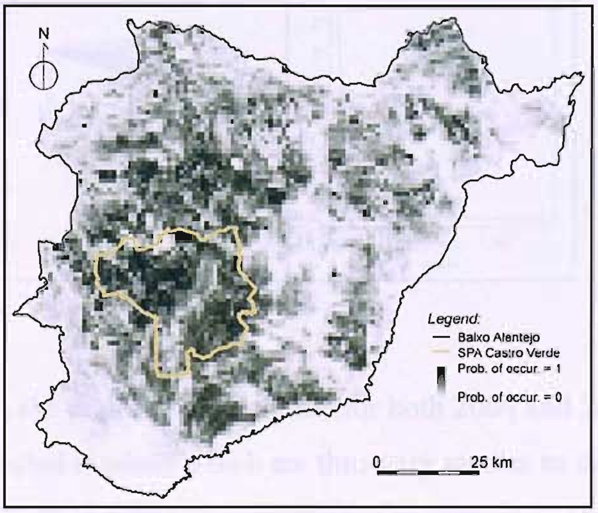
In the selected regional scale Little Bustard (*Tettet*) models (Table 5.4), *Dry* was by far the most contributing variable, with a positive response to areas with lower *Dry* values (Figure 5.12). Other contributing variables were *Wet*, *Summer*, *May*, *Topov10*, *Waterdist* and *Roaddist*.

Figure 5.12 - Regional scale model partial plots for Little Bustard (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



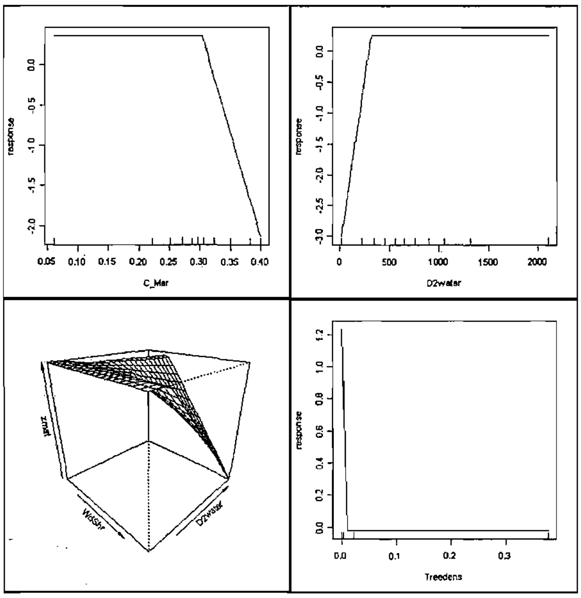
The predicted occurrence patterns of Little Bustard in the Baixo Alentejo show an avoidance of the hilly areas, as well as the steep areas of the river valleys (Figure 5.13). It is also possible to observe a preference of the species for the Castro Verde plains, with a mean predicted probability value within the SPA of 209.8 % of that for the Baixo Alentejo (see Figure 5.5).

Figure 5.13 - Predicted probabilities of occurrence of Little Bustard in the Baixo Alentejo



The selected landscape scale models never achieved predicting performance (Table 5.6). Nevertheless, these models had high (over 50 %) drop contribution of the variable *C_Mar*, also with considerable contributions from *D2water*, *Treedens* and *WdShr* (Table 5.7). Inspection of the species responses to the predictor variables suggests they avoid areas with high cereal coverage, close to water bodies, and areas with tree density over 0.025 – roughly equivalent to one 5 x 5 m pixel classed as "*tree*" within each 30 x 30 m pixel (Figure 5.14). From all studied species, this was the only one that selected the *Treedens* (instead of *D2tree*) at the pre-modelling variable reduction phase. Additionally, all selected models included an interaction between *WdShr* and *D2water*, by avoiding areas with high values on both variables. Minor contributing variables in the selected models were *Fallow*, *D2built* and *BS_Mar*. No variables describing the terrain were included in any of the selected landscape scale models for this species.

Figure 5.14 - Landscape scale model partial plots for Little Bustard (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)

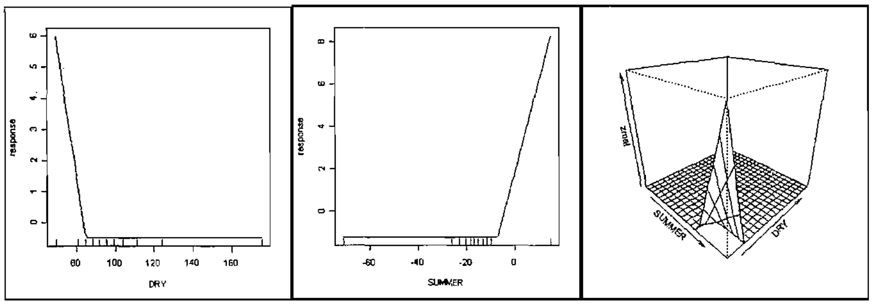


In the MS analysis, the regional predictions (for both 2004 and 2006) were not included in the selected models, which are thus very similar to the respective landscape models, both in terms of performance and structure (Table 5.9; Table

5.10). This way, the predicted regional distribution patterns for this species did not improve the fitting of those observed at the landscape scale within the Castro Verde study area.

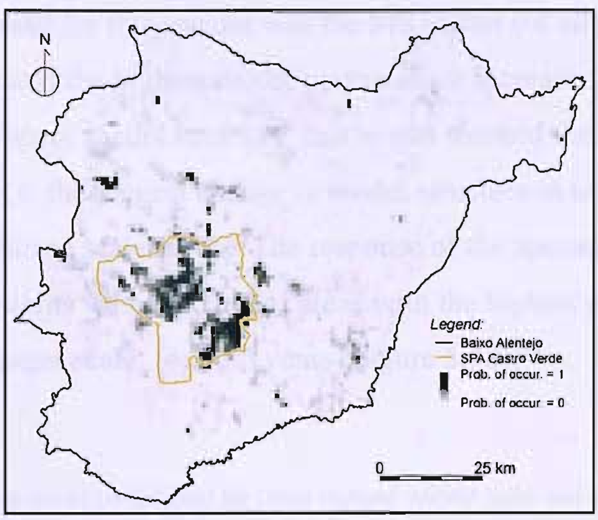
At the regional scale, the selected Great Bustard (*Otitar*) models (Table 5.4) had greater contribution from variable *Dry* (Table 5.5), being the areas with the lowest values the most favourable ones (Figure 5.15). The second most contributing variable was *Summer*, the species selecting those areas with little decrease in NDVI during the Summer months (June to September). In one of the selected models (R50), however, an interaction between these two variables was fitted, with a positive response to areas with simultaneously low *Dry* and high *Summer* values. Other minor contributing variables included *Dec*, *Wet*, *May*, *Spring* and *Roaddist*. No terrain variables were used in the selected models.

Figure 5.15 - Regional scale model partial plots for Great Bustard (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



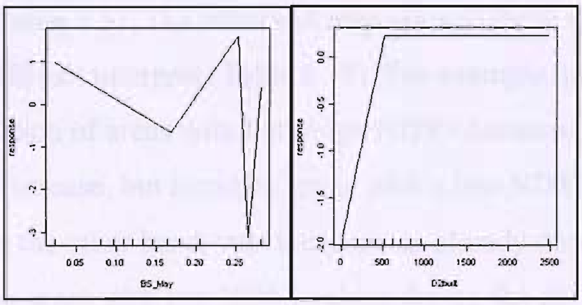
The predicted distribution of Great Bustard in the Baixo Alentejo was very sparse and restricted to two main areas within the Castro Verde SPA plus a few small nuclei elsewhere (Table 5.16). Indeed, the mean prediction value for Castro Verde is 515.6 % of that for the whole Baixo Alentejo, this showing a sharply marked preference of this species for the Castro Verde pseudo-steppes (see Figure 5.5).

Figure 5.16 - Predicted probabilities of occurrence of Great Bustard in the Baixo Alentejo



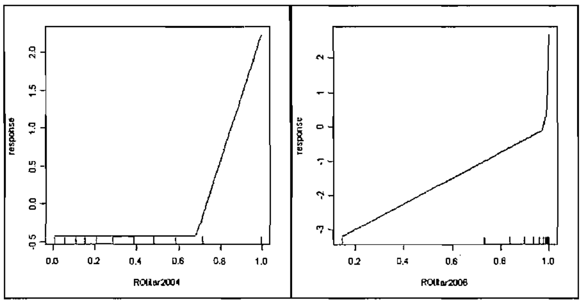
The Great Bustard landscape models (Table 5.6) used mostly the *BS_May* variable (Table 5.7), even though with a response of difficult interpretation (Figure 5.17). The second most contributing variable was *D2built*, with the species avoiding areas closer to 500 m to the nearest built-up structure. Minor contributing variables were *Fallow*, *D2water*, *WdShr*, *DV_May*, *Slope* and *C_Mar*. These models, however, never achieved predicting performance, so no predictions were generated for this species at this scale.

Figure 5.17 - Landscape scale model partial plots for Great Bustard (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



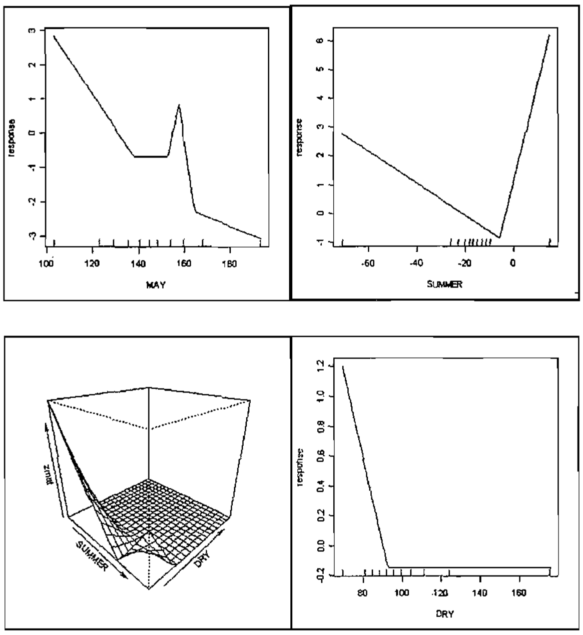
Both MS models for this species (M2004 and M2006) had high contributions of the regional scale prediction variables (Table 5.10), with considerable improvements in the fitting of the landscape bird data, even though still resulting in weak performances of the selected models (Table 5.9). In particular the selected M2006 model for this species was the MS model (of all selected for all species) which showed the highest model performance increase (ΔAUCcv = + 0.080). Also, in terms of model structure, this model showed the lowest τ value of all MS models, i.e. the biggest change in model structure in comparison with the respective landscape scale model. The response of the species to the regional scale prediction patterns was selection of areas with the highest probability of occurrence at the larger scale, for both years (Figure 5.18).

Figure 5.18 - Multi-scale model partial plots for Great Bustard: M2004 (left); and M2006 (right) (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



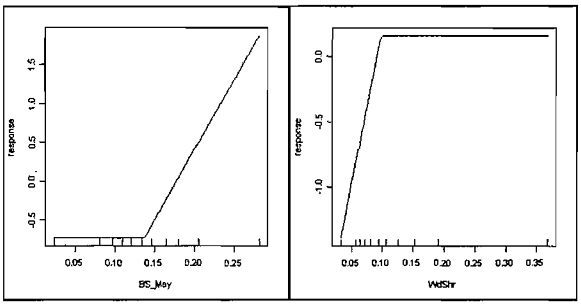
The Stone Curlew (*Buroed*) models never achieved "good" performance, and thus its probability patterns were not predicted for this species. Nevertheless, the selected regional models (Table 5.4) were mostly based on the variables *May*, *Summer* and *Dry* (Table 5.5).. The observed responses to these variables were, however, mostly difficult interpret (Table 5.19). For example, the Stone Curlew data showed a selection of areas with both high NDVI decrease in the summer months or with an increase, but avoiding areas with a low NDVI decrease. The response to *Dry*, on the other hand, was the same as already observed for other species, of selecting areas with low NDVI values during the dry months.

Figure 5.19 - Regional scale model partial plots for Stone Curlew (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



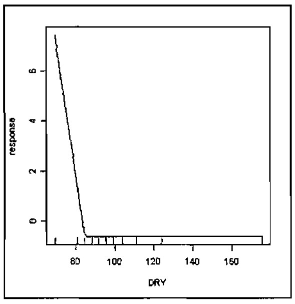
The selected landscape models (Table 5.6) were all built on only two variables (*BS_May* and *WdShr*) although with much higher contribution of *BS_May*, therefore never using any terrain or disturbance related variables (Table 5.7). Nevertheless, within these data, Stone Curlew showed a preference for areas with high proportion of bare soil in May, although avoiding areas with little or no woodlands / shrubs (Figure 5.20).

Figure 5.20 - Landscape scale model partial plots for Stone Curlew (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



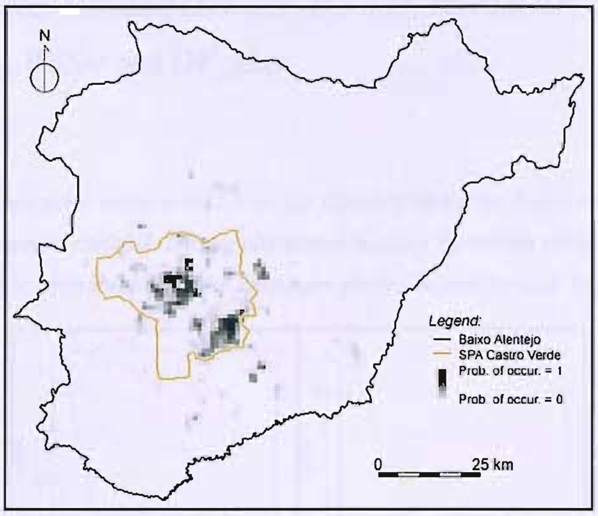
Black-bellied Sandgrouse (*Pteori*) data were only collected at the regional scale, so no landscape and MS models were run. The selected regional scale models (Table 5.4) had very high contributions of the *Dry* variable, with over 70 % (on average) of deviance explained loss when this variable was dropped (Table 5.5). As for other steppe bird species, the Black-bellied Sandgrouse selected areas with low *Dry* values (Figure 5.21). Other minor contributing variables in the selected models were *Dec*, *Waterdist*, *Wet* and *Spring*. Thus, no variables describing terrain contributed for the selected models.

Figure 5.21 - Regional scale model partial plots for Black-bellied Sandgrouse (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



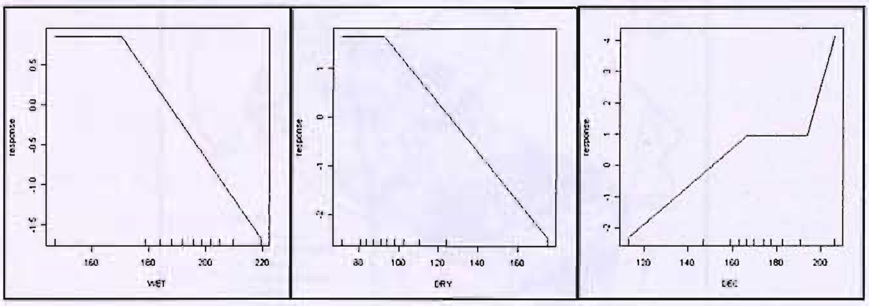
The predicted distribution of this species in the Baixo Alentejo is mostly confined to pseudo-steppe areas within the Castro Verde SPA (Figure 5.22), with mean prediction in this area constituting 632.0 % of that for the Baixo Alentejo (see Figure 5.5).

Figure 5.22 - Predicted probabilities of occurrence of Black-bellied Sandgrouse in the Baixo Alentejo



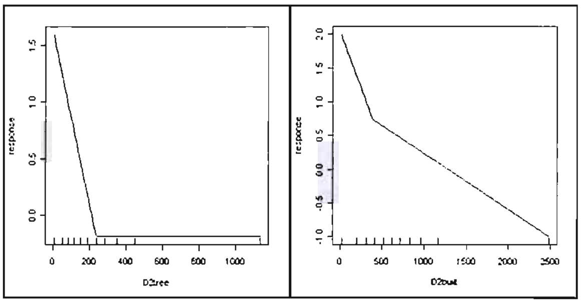
The regional scale models for the *Galerida* larks (*Galsp*) did not achieve predicting performance (Table 5.4). However these models were built on three main variables: *Wet*, *Dry* and *Dec* (Table 5.5). In them, the species avoided areas with high values on the *Wet* and *Dry* and low values on the *Dec* variable (Figure 5.23). Other variables included in these models were *Winter*, *Summer*, *Waterdist* and *Topov10*.

Figure 5.23 - Regional scale model partial plots for Galerida *larks (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)*



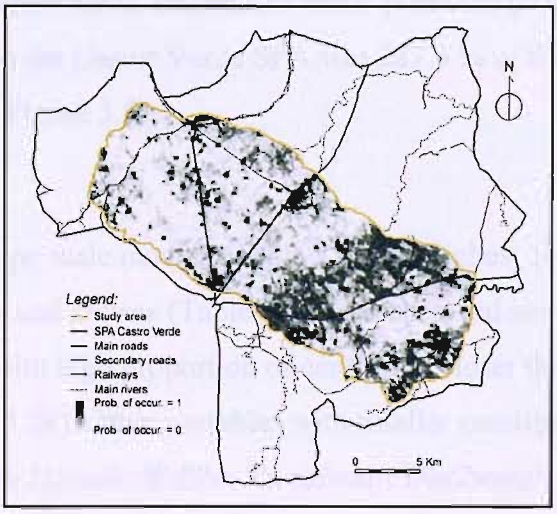
The landscape scale models, on the other hand, achieved the highest model performances of all landscape models for all species (Table 5.6). The models had large contributions of the disturbance variables *D2built* and *D2tree* (Table 5.7), with the species occurring mostly in areas closer than 250 m to the nearest tree and avoiding areas far from built-up structures (Figure 5.24). Other less contributing variables included *D2water*, *D2road*, *Terrainvar*, *C_Mar*, *C_May*, *BS_Mar*, *BS_May*, *WdShr* and *DV_May*.

Figure 5.24 - Landscape scale model partial plots for Galerida *larks (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)*



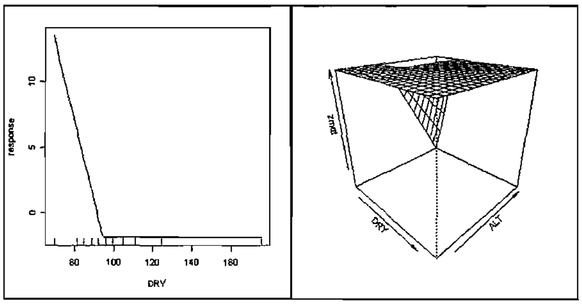
The predicted distribution of these species in the Castro Verde study area showed their occurrence mostly around the valleys of the two main rivers, the extreme SE of the area, plus some small circular nuclei around farmhouses and local villages (Figure 5.25).

Figure 5.25 - Predicted probabilities of occurrence of Galerida *larks in the Castro Verde SPA*



The selected regional models (Table 5.4) of Calandra Lark (*Melcal*) had high contributions of the variable *Dry* (Table 5.5). On these models, Calandra Lark preferred areas with low values of this variable (Figure 5.26). In one of the selected models (R60) an interaction was fitted between this variable and *Alt* (the second most contributing variable), even though this is of difficult interpretation and as it was not fitted consistently on all selected models it may possibly be a statistical artefact, rather than a real species response. Other variables with minor contributions in these models were *Summer*, *Dec*, *Urbandist*, *Wet*, *Topov10* and *Waterdist*.

Figure 5.26 - Regional scale model partial plots for Calandra Lark (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



The predicted distribution patterns of this species show it mainly occurs inside the Castro Verde SPA and in an area south of it, with some additional small nuclei elsewhere (Figure 5.27). Indeed, the mean predicted probability values for Calandra Lark within the Castor Verde SPA was 287.4 % of that for the whole Baixo Alentejo (see Figure 5.5).

The selected landscape scale models (Table 5.6) had highest contributions from the variables *C_Mar* and *D2tree* (Table 5.7). The observed responses were of avoidance of areas with high proportion of cereal and closer than 250 m to the nearest tree (Figure 5.28). Other variables with smaller contributions in the selected models were *D2built*, *WdShr*, *Terrainvar*, *Dist2water*, *Dist2road*, *BS_May*, *Fallow*, *DV_May* and *C_May*.

Figure 5.27 - Predicted probabilities of occurrence of Calandra Lark in the Baixo Alentejo (left) and in the Castro Verde SPA (right)

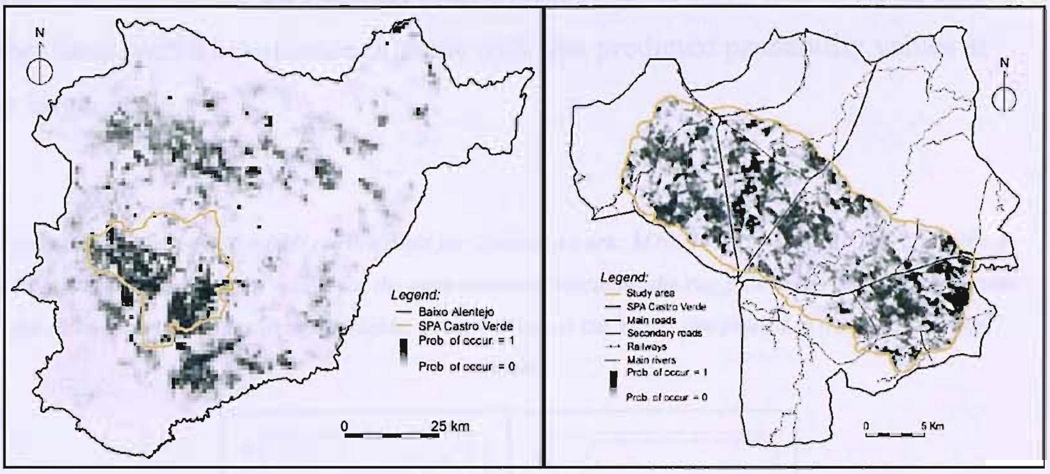
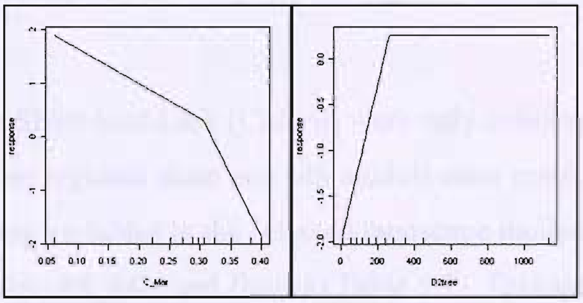


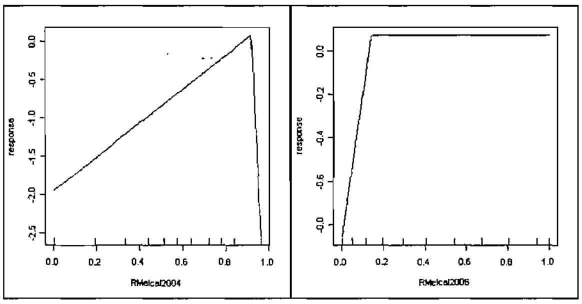
Figure 5.28 - Landscape scale model partial plots for Calandra Lark (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



The predicted probabilities of Calandra Lark within the Castro Verde study area show a generalised distribution throughout the area, although with a patchy pattern, and mostly absent in the areas close to the main rivers – which was also observed in the regional model predicted distributions (Figure 5.27).

Both MS model selected the regional scale predictions as predictor variables, even though with small contributions, and respective little change in model performance and structure (Table 5.9; Table 5.10). The observed responses of the species to the regional predictions for 2004 (M2004) were of increasing landscape scale probabilities with the increasing regional scale probabilities up to an optimal value of ca. 0.90, and a sharp decline on landscape scale probability of occurrence for areas with regional probabilities close to 1 (Figure 5.29). The observed response to the regional scale predictions for 2006 (M2006), on the other hand, was of avoidance of areas with low predicted probability values at the larger scale.

Figure 5.29 - Multi-scale model partial plots for Calandra Lark: M2004 (left); and M2006 (right) (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



Occurrence data for Short-toed Lark (*Calbra*) were only collected at the landscape scale, so no regional scale and MS models were run for this species. The most contributing variables in the selected landscape models (Table 5.6) were *C_Mar*, *BS_May*, *BS_Mar* and *D2tree* (Table 5.7). This species selected areas with intermediate values for *C_Mar* (which is of difficult interpretation) and generally with a high proportion of base soil, while avoiding areas closer than 150 m to the nearest tree (Figure 5.30). Minor contributing variables were *C_May* and *D2built*, hence without the inclusion of terrain-related variables.

The predicted occurrence patterns of Short-toed Lark show a scarce distribution, restricted mostly to specific landscape patches, namely those which remained ploughed during the breeding season (Figure 5.31).

Figure 5.30 - Landscape scale model partial plots for Short-toed Lark (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)

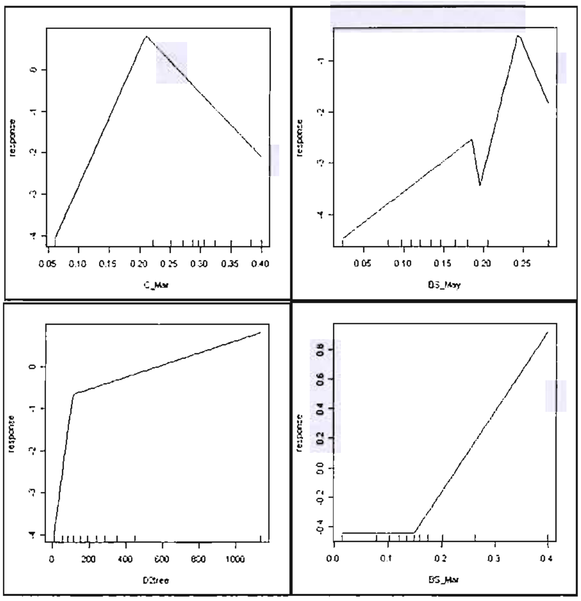
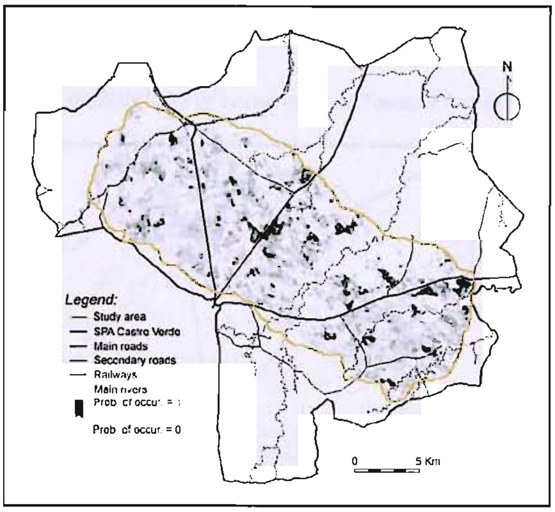
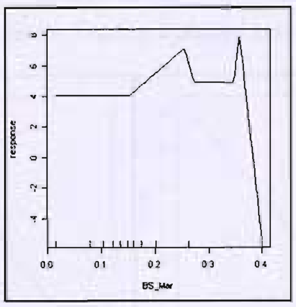


Figure 5.31 - Predicted probabilities of occurrence of Short-toed Lark in the Castro Verde SPA



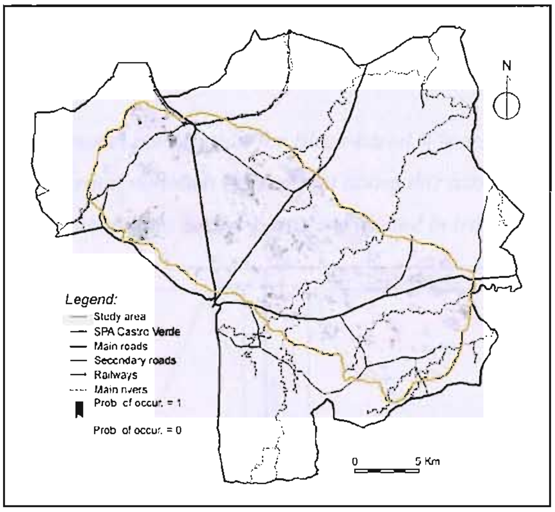
Tawny Pipit (*Antcam*) occurrence data were only collected at the landscape scale, so no regional scale and MS models were run for this species. The selected landscape scale models (Table 5.6) were only built on two predictor variables: the most contributing *BS_Mar*; and the least contributing *D2water* (Table 5.7). Hence, no terrain variables were used to model this species' distribution. The fitted responses indicate that the species selects areas with a relatively high proportion of bare soil (in March), although it seems to avoid areas of full bare soil coverage (Figure 5.32).

Figure 5.32 - Landscape scale model partial plots for Tawny Pipit (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



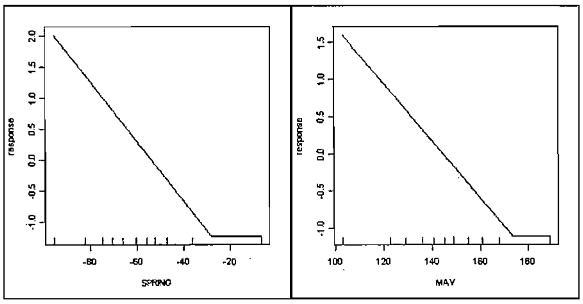
The predictions of occurrence of Tawny Pipit in the Castro Verde study area shows it to be well distributed throughout the area, although probably with low densities – low values of probability of occurrence (Figure 5.33).

Figure 5.33 - Predicted probabilities of occurrence of Tawny Pipit in the Castro Verde SPA



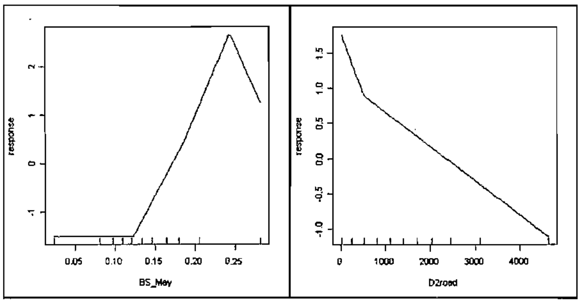
In the case of Black-eared Wheatear (*Oenhis*), the achieved performances of the selected regional models of did not allow for prediction of the species distributions in the Baixo Alentejo (Table 5.4). These models were fit using only two variables: *Spring* (the most contributing) and *May*, so always excluding variables descriptive of terrain and disturbance (Table 5.5). According to the response curves plotted (Figure 5.34), this species selected areas with low values in the two most contributing variables.

Figure 5.34 - Regional scale model partial plots for Black-eared Wheatear (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



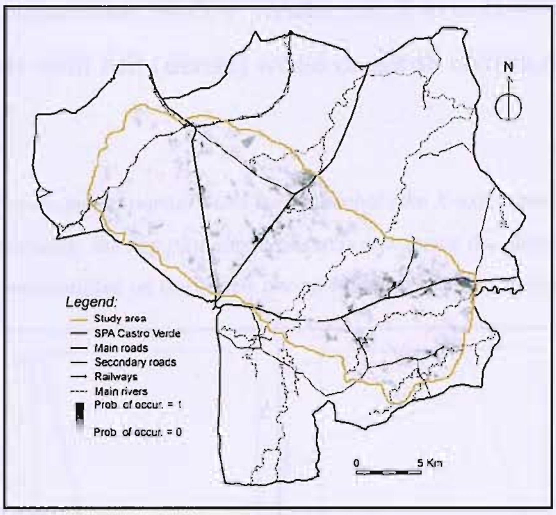
The selected landscape scale models (Table 5.6) had higher contributions from variables *BS_May* and *D2road*, but also with lower contribution of *Fallow*, *WdShr*, *Slope*, *D2built* and *C_Mar* (Table 5.7). The species response to the main contributing variables was of selection of areas with a generally high proportion of bare soil and close to roads (Figure 5.35).

Figure 5.35 - Landscape scale model partial plots for Black-eared Wheatear (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



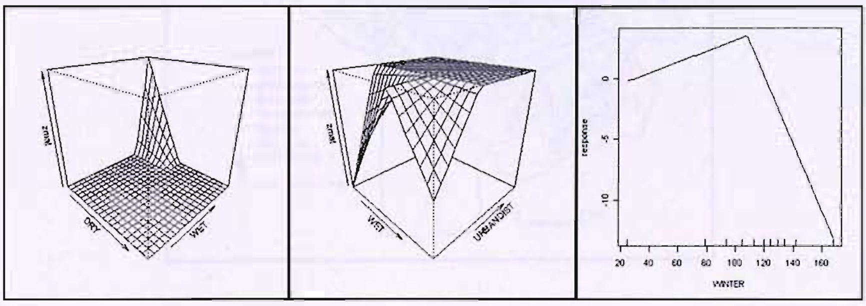
According to the predicted probabilities of occurrence, Black-eared Wheatear is mostly distributed in areas of bare soil (such as ploughed fields) not far from the main roads (Figure 5.36). However, the relatively low predicted probability values suggest it occurs in low densities.

Figure 5.36 - Predicted probabilities of occurrence of Black-eared Wheatear in the Castro Verde SPA



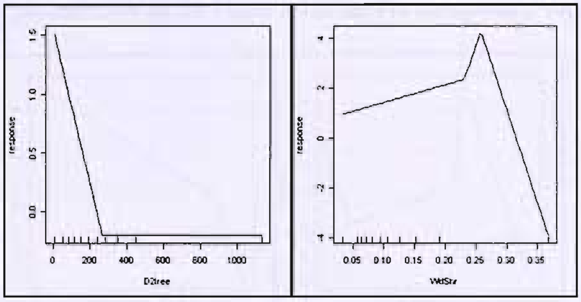
The selected regional models of Stonechat (*Saxtor)* did not achieve predicting performance (Table 5.4). However, they had higher contributions of the *Wet* and *Winter* variables, although with small contributions of *Waterdist*, *Summer*, *Topov10*, *Dry*, *May*, *Dec*, *Alt*, *Roaddist* and *Urbandist* (Table 5.5). The most contributing variable ("*Wet*") was however always selected in interaction with other variables, like *Dry* and *Urbandist*, making difficult the interpretation of the species responses (Figure 5.37). On the other hand, areas with high values of *Winter* were avoided by this species.

Figure 5.37 - Regional scale model partial plots for Stonechat (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



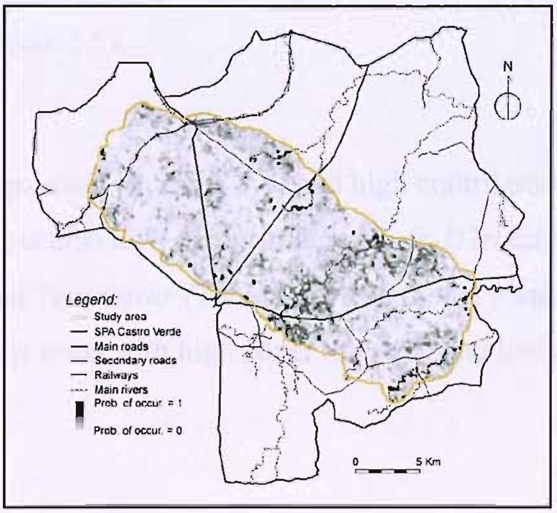
The selected Stonechat landscape models (Table 5.6) had higher contributions of the variable *D2tree* and *WdShr* (Table 5.7). The model partial plots (Figure 5.38) indicate that Stonechat avoids areas further than 250 m from the nearest tree, and selects areas with intermediate *WdShr* values, i.e. it avoids areas without trees or shrubs, but also areas with full (dense) wood or shrub coverage.

Figure 5.38 - Landscape scale model partial plots for Stonechat (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



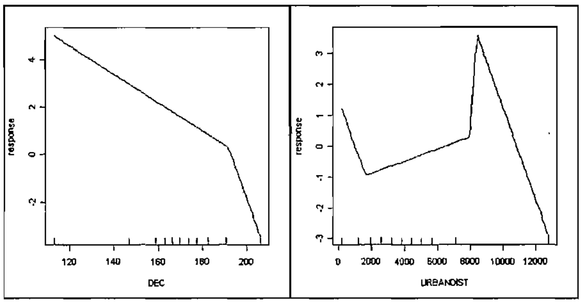
The predicted probabilities of occurrence show a generalised distribution throughout the area, although in greater densities (higher probability) close to the two main rivers in the area (Figure 5.39).

Figure 5.39 - Predicted probabilities of occurrence of Stonechat in the Castro Verde SPA



At the regional scale, the Zitting Cisticola (*Cisjun*) models (Table 5.4) had greater contributions from *Dec* and *Urbandist*, but also minor contributions from *May*, *Wet*, *Winter*, *Summer*, *Alt* and *Topov10* (Table 5.5). The species preferentially selected areas with low NDVI values in the month of December (Figure 5.40). Its response to *Urbandist*, however consistent among all selected models was difficult to interpret, showing a bimodal effect with optima at 0 and 9000 m, while avoiding areas at 2000m from the nearest urban / built-up area.

Figure 5.40 - Regional scale model partial plots for Zitting Cisticola (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



According to the predicted probabilities of occurrence, Zitting Cisticola occurs throughout the Baixo Alentejo, although avoiding some of the hilly or steep areas (of the Serra do Caldeirão and in the Guadiana valley) as well as the S and E edges of the area. The species did not show particular preference for the Castro Verde SPA, and the mean prediction value for this area is 122.9 % of that for the whole region (see Figure 5.5).

The selected landscape models (Table 5.6) had high contributions from variables *C_Mar* and *Fallow*, but also lower contributions from *D2road*, *DV_May*, *WdShr*, *BS_May*, *BS_Mar* and *Terrainvar* (Table 5.7). The model partial plots indicate that the species selects areas with high cover of cereal and low of fallow (Figure 5.42).

Figure 5.41 - Predicted probabilities of occurrence of Zitting Cisticola in the Baixo Alentejo (left) and in the Castro Verde SPA (right)

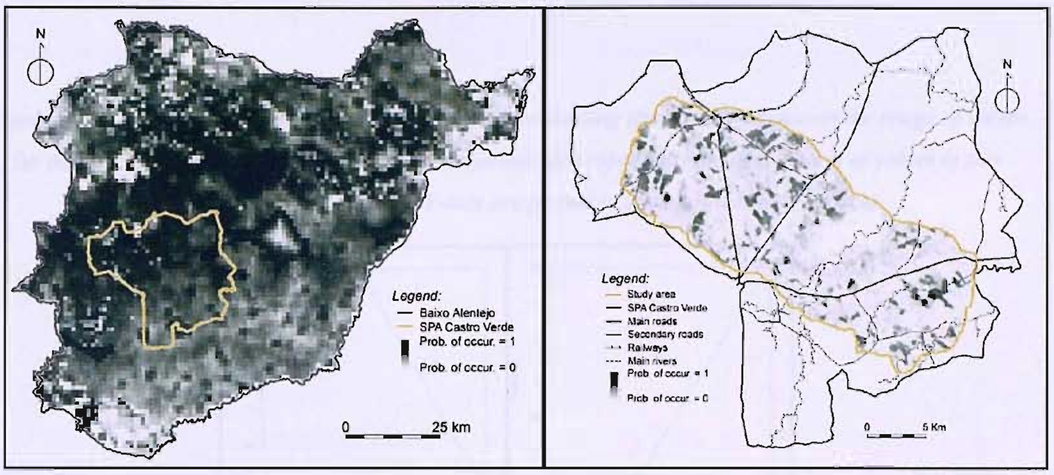
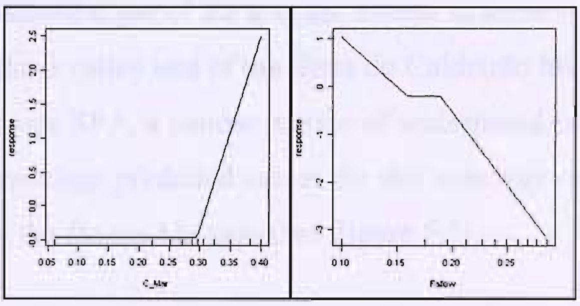


Figure 5.42 - Landscape scale model partial plots for Zitting Cisticola (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)

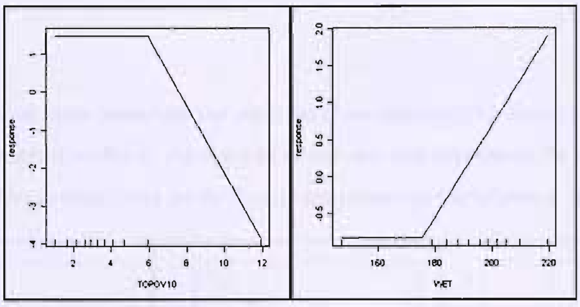


The selected MS model for this species did not use the regional predictions as landscape predictor variables, which resulted in roughly the same models as at the landscape scale, with little change in model performance or structure (Table 5.9; Table 5.10). From this, and as for Little Bustard, it can only be concluded that the predicted regional distribution patterns for this species do not help to describe those observed at the landscape scale within the Castro Verde study area.

Finally, the selected regional Corn Bunting (*Milcal*) models (Table 5.4) showed high contributions from the *Topov10* and *Wet* variables, but also lower contributions from *Dec*, *Summer*, *Dry* and *Alt*, hence with no contributions from disturbance-related variables (Table 5.5). The model partial plots show that Corn

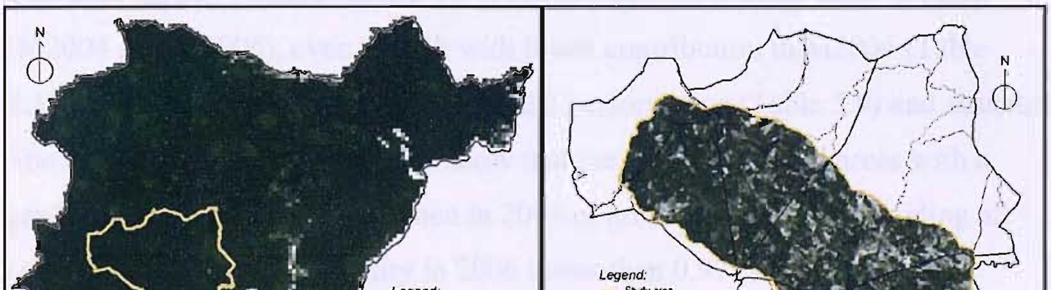
Bunting avoided areas with high topographic variability, as well as with low NDVI values between the months of January to April (Figure 5.43).

Figure 5.43 - Regional scale model partial plots for Corn Bunting (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



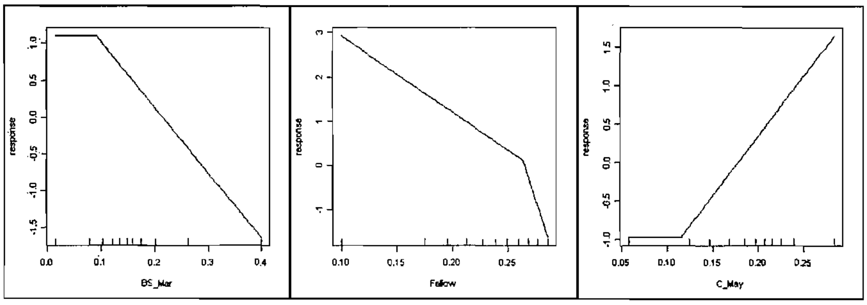
The predicted probabilities of occurrence of Corn Bunting in the Baixo Alentejo showed widespread distribution of the species, except in some small areas such as parts of the Guadiana valley and of the Serra do Caldeirão hills (Figure 5.44). Within the Castro Verde SPA, a similar pattern of widespread occurrence was observed – indeed the mean predicted values for this area was very similar (105.5 %) to that for the Baixo Alentejo (see Figure 5.5).

Figure 5.44 - Predicted probabilities of occurrence of Corn Bunting in the Baixo Alentejo (left) and in the Castro Verde SPA (right)



The variables most contributing to the respective selected landscape scale models (Table 5.6) were *BS_May*, *Fallow* and *C_May*. Least contributing variables in these models were *DV_May*, *BS_May*, *WdShr*, *D2built* and *C_Mar*, thus not incorporating variables describing terrain (Table 5.7). Corn Bunting seemed to prefer areas without bare soil, and with low fallow and high cereal cover (Figure 5.45).

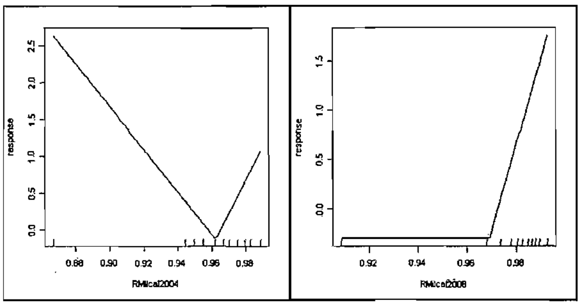
Figure 5.45 - Landscape scale model partial plots for Corn Bunting (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



The predicted patterns of occurrence at this scale also showed a generalised distribution of the species throughout the Castro Verde study area, although avoiding some particular land patches as e.g. ploughed fields (Figure 5.44).

The selected MS models did use the regional scale predictions on both cases (M2004 and M2006), even though with lower contribution in M2004 (Table 5.10), this reflected in terms of both model performance (Table 5.9) and structure. The partial plots for these models show that the species avoided areas with a predicted probability of occurrence in 2004 of around 0.96, while avoiding all areas with predicted probability in 2006 lower than 0.97 (Figure 5.46).

Figure 5.46 - Multi-scale model partial plots for Corn Bunting: M2004 (left); and M2006 (right) (the X-axis represents the range of values for the environmental variable; the rug plot above this axis represents the distribution of values in this variable; probabilities on the Y-axis are plotted in transformed 'logit' space)



5.4. *Discussion*

The modelling framework used in this chapter, by replicating the predictor datasets into three samples (for each species) with different correlation thresholds between variables, together with the implementation of an (exhaustive) grid search for the selection of the respective model parameters, allowed the optimisation of the model fitting for each species. The use of three models with different predictor subsets (with varying correlation thresholds), for each species can be considered as an extension of the grid search approach. Data collinearity, on the other hand, is usually referred to as a source of uncertainty in the interpretation of these models (Friedman 1991; Morlini 2006). However, by averaging the drop contributions of the predictor variables across the three model replicates (for each species) it was possible to identify the predictors that consistently influenced the observed occurrence patterns, across different levels of data collinearity. Also, model averaging and other multi-model inference approaches are considered to be superior to single-model ones (Buckland *et al.* 1997; MacNally 2000; Burnham & Anderson 2004).

The hierarchical (top-down) MS modelling approach applied here, with the incorporation of larger-scale predictions as predictor variables in the finer-scale models, follows work done by Pearson *et al.* (2004), which used it for predicting four plant species in Britain with ANN models. Other MS approaches include the use of large-scale occurrence data instead of model predictions (Araújo *et al.* 2005; McPherson *et al.* 2006) or the non-hierarchically inclusion of variables collected at different grains in a single model (Coreau & Martin 2007).

The prediction of the regional scale models over 2005/6 vegetation data (models M2006) assumes model transferability between the two years, and within the Castro Verde study area. In fact, by describing the annual vegetation phenological cycle, these variables can be considered to be proximal and the respective models, therefore contain a certain degree of generality (Austin 2002). Additionally, all fitted responses to the 2006 regional predictions (for Great Bustard, Calandra Lark and Corn Bunting) were consistent with the expected, with increasing probabilities of occurrence at the landscape scale with the increasing probabilities at the regional scale (Figure 5.18; Figure 5.29; Figure 5.46). For the species whose MS models did not select this variable, however, no conclusions should be taken in respect to scale effects on habitat selection or on the respective distribution patterns. On the other hand, on models M2004, responses showing species avoiding (at the landscape scale) the areas with high regional probabilities in 2004 (Figure 5.7; Figure 5.29), or other more complex responses (Figure 5.46) may be related with the existing agricultural crop rotation system. Indeed, the most suitable areas for a certain species in a particular year most likely will have an opposite land use (e.g. fallow to cereal) two years later. Nevertheless, due to the uncertainty associated with the transferability of these models, no conclusions should be taken from those species whose models did not select the regional predictions in the landscape models. The Great Bustard, on the other hand, was positively associated with regional predictions on both years, which probably relates to site fidelity, observed on this species (Alonso *et al.* 2000; Martín *et al.* 2002).

The use of image soft classification outputs in the landscape scale models is not without implications for model interpretation. Unlike hard classification outputs where a high value of a certain class would be directly related to the proportion of cover of the circular-plot by the respective class, the same may not be true for predictors derived from soft classes. For example, a medium value of a certain soft class, i.e. an average intermediate probability of membership of the pixels (within the circular-plot) to the respective class, may be the result of the process above described (as for hard classes), but it could also mean that those pixels belong to another class which is spectrally close to that described (Foody 1999). For this reason, some care must be taken in interpreting the species responses along the gradients represented by these predictors. In fact this could partially explain the complex response curves found in some of the models (e.g. *BS_May* in Figure 5.17 or May in Figure 5.19).

Also, the use of high-resolution LiDAR data considerably improved the models for some species (Table 5.8). In particular, on three occasions, LiDAR-derived variables were the most contributing in the respective models ("*Slope*" for Red-legged Partridge and *D2tree* for *Galerida* larks and Stonechat), and relatively important variables on five other cases (*Treedens* for Little Bustard, *D2built* for Great Bustard and *Galerida* larks, and *D2tree* for Calandra Lark and Short-toed Lark). This corroborates the findings of other authors, which have unveiled the potential of LiDAR data for generating habitat descriptor variables in agricultural and woodland areas (Mason *et al.* 2003; Bradbury *et al.* 2005).

This study greatly improved existing knowledge about steppe bird habitat selection in South Portugal, and in particularly in the Castro Verde pseudo-steppes. This study greatly improved existing knowledge about steppe bird habitat selection in South Portugal, and in particularly in the Castro Verde pseudo-steppes. In terms of their habitat preferences, the studied species could be grouped in four main groups. A first group "Dry-Fallow" would include species which selected, at the regional scale, areas with little vegetation during the dry summer months – low *Dry* values, mostly associated with the pseudo-steppe landscapes – and/or, at the landscape scale, areas of fallow (usually expressed in the models as avoiding the second most common habitat, the cereal fields): Little Bustard and Calandra Lark. A second group "Dry-Ploughed" would be constituted by species also selecting little vegetated areas in the summer at the regional scale, and/or areas of bare soil / ploughed fields at the landscape scale: Stone Curlew, Black-bellied Sandgrouse, Short-toed Lark, Tawny Pipit and Black-eared Wheatear. Species which selected, at the regional scale, areas with vigorous vegetation during the wet winter months (high values of *Wet*, mostly associated to winter cereal crops) and/or high proportion of cereal (*C_Mar* or *C_May*), at the landscape scale, would constitute a third group "Wet-Cereal": Quail, Zitting Cisticola and Corn Bunting. A forth group "Others" would include species which are not associated with any of the previous, but rather mostly to other features in the landscape, such as trees, shrubs or built-up areas: Red-legged Partridge, *Galerida* larks and Stonechat. Two species, however, were difficult to fit into these groups. Montagu's Harrier, which at the regional scale selected dry (pseudo-steppe) areas, at the landscape scale was associated to cereal fields (which could theoretically constitute a fifth group "Dry-Cereal"). Great Bustard, on the other hand, while clearly selecting dry areas (regionally), at the landscape level it was difficult to interpret its habitat preferences (besides its avoidance of built-up structures). Also, the *Galerida* larks, while included in the forth group of species, at the regional scale also selected the (pseudo-steppe) dry areas. These four main groups of species mostly agree with the findings of Moreira *et al.* (2007) (see Appendix A.1), with some differences. For example, Montagu's Harrier had been included with the cereal-associated species by these authors, even though positive associations with fallows with shrubs had also been found. Also, in the referred study, no habitat associations had been found for Stone Curlew, while other studies had associated its presence in fallows to shrub occurrence (Moreira 1999) and its abundance to ploughed fields (Delgado & Moreira 2000). In the present study it was possible to identify its preference to ploughed fields, while also to areas of shrubs, nevertheless placing it within the group "Dry-Ploughed". Additionally, the stronger fit of the Little Bustard occurrence data to tree density as opposed to the distance to the nearest tree, also agrees with the results of Moreira *et al.* (2007) (see Appendix A.1) where the species responded to the abundance of holm oak 'montados' within the circular-plot instead of its presence.

Besides the improvement (and slight rearrangement) of the grouping of the steppe bird community of Castro Verde according to their habitat preferences, other species-habitat associations were found, some of which were not known before. For example, while the avoidance of Calandra Lark to afforestation has been investigated in a recent study (Reino *et al. in press*), the disturbance effect of single trees in the landscape has been described and quantified for the first time in the present study, the bird avoiding areas closer than 250 m to the nearest tree. Additionally, a similar although smaller disturbance effect (of individual trees in the steppe landscape) on Short-toed Lark was also found and quantified: areas closer to 150 m to the nearest tree being avoided. Similarly, it was possible to quantify the disturbance effect of built-up structures for Great Bustards: areas closer than 500 m to the nearest house or built-up structure being avoided.

Additionally to these findings, it was possible to infer some aspects of habitat selection relating to scale effects by investigating species-environment responses at two different spatial scales. By comparison of the achieved performances of the models, it is possible to observe that generally species with larger body sizes tended to be better modelled at the large spatial scale. From another perspective, only the models for the (small sized) passerine birds (from *Galerida* larks to Corn Bunting on Table 1.1) achieved predictive performance at the landscape scale. In fact, only three species (*Galerida* larks, Zitting Cisticola and Corn Bunting) were fitted at both scales with predictive performance. This opens the way for future macroecological research (Blackburn & Gaston 2002), such as e.g. investigating the ideal grain size to explain the various species distributions.

However, with the present study it was possible to decouple some scale effects, as for example, for the Montagu's Harrier. In the work by Moreira *et al.* (2007) (see Appendix A.1), this species was positively associated with both cereal fields and fallows with scattered shrubs. On the other hand, in the obtained SDMs, this species seemed to select areas of pseudo-steppe (low *Dry*) at the larger (regional) scale, while selecting areas of cereal at the finer (landscape) level.

In the MS analysis, the regional scale predictions only showed clear improvements in the fitting of the landscape scale responses for one species, the Great Bustard. From this result it is possible to interpret that its large scale distribution patterns, which result from processes functioning at a regional scale, are capable of explaining its landscape scale occurrences. It may be then assumed that this species responds to processes that occur at a spatial scale larger than that of the landscape scale study. From the remaining five species modelled with the MS approach, no conclusions can be drawn, as the regional scale patterns either were not selected by the respective MS models, or the resulting model performance did not change greatly.

Finally, the distribution patterns of the studied birds were predicted for the respective study areas (when the model performance allowed for prediction). In the case where predictions were made at both scales, it was possible to observe the concordance of both predicted distributions within the Castro Verde SPA. Furthermore, the species occurrence patterns in Castro Verde, which were originally described (recurring to spatial interpolation of the presence points) by Moreira *et al.* (2007) (see Appendix A.1) were further improved by the addition of habitat contextual information within the architecture of the present SDMs.

6. Synthesis and general discussion

[illegible]

[illegible]

[illegible] distribution of [illegible] in the study region [illegible] knowledge about [illegible] species [illegible]

6.1. Synthesis

The work here presented had as its main goal the improvement the knowledge and understanding of the occurrence patterns and habitat preferences of the steppe bird community of southern Portugal, while accounting for the effects of spatial scale.

A second goal of this work was a methodological one. It aimed at making the best use of robust statistical analysis methods, commonly called species distribution models (SDMs), in order to achieve the first goal. In other words, it aimed at improving or optimising off-the-shelf tools for answering the particular ecological question for which was proposed.

Special attention was paid to the data used as input in the models, the derivation of both the response and predictor variables, and the effects of data quality in terms of model performance, prediction and interpretation. In particular, the effects of (species) data sampling bias (see Chapter 2), and of the use of different processing methods for information extraction from remote sensing imagery – the main used source of environmental descriptive data used (see Chapters 3 and 4) – were explored. Finally, the effects of spatial scale were incorporated, by performing the analyses at two different scales, including a hierarchical multi-scale model integration (see Chapter 5).

Overall, the proposed objectives were achieved. In terms of the ecology and distribution of steppe birds in the study region, this study generated new knowledge about some species-habitat associations (as for Calandra Lark and Short-toed Lark), decoupled some spatial scale effects (e.g. Montagu's Harrier and Great Bustard), and further described the resulting occurrence patterns of most studied species.

Methodologically, it made significant progress in the understanding of the functioning of SDMs and their sensitivity to the data characteristics. From this understanding, it was possible to optimise these models in order to answer the research question. This optimisation was done through the improvement of the input data quality, by using advanced data processing techniques (such as the use of SVM soft classification) and through a robust methodological approach (like the use of multiple predictor subsets and the selection of the model parameters). These achievements, however, were only possible with the use of high quality environmental datasets (such as the LiDAR data) and by extensive and carefully planned fieldwork campaigns, capable of collecting large (and statistically balanced) species locational datasets.

Nevertheless, it is considered the added-knowledge about the steppe birds in the region justifies the long data processing involved and the complex methodological set-up. Furthermore, it is recommended that this added-knowledge is translated into new (or incorporated into existing) management recommendations for a more effective conservation of these species and their habitats. There are several issues, however, relevant to the work presented, which should be further discussed.

6.2. Discussion

In this study, species-environment associations were investigated by using species occurrence data. While this is a common practice (Martínez 1994; Carol *et al.* 1999; Suárez-Seoane *et al.* 2002a; Moreira *et al.* 2004; Rosalino *et al.* 2008), it is also clear that the incorporation of species abundance data could provide further insights into their habitat preferences (Ralph 1985; Leitão & Costa 2001; Silva *et al.* 2007). While there are clear relationships between distribution and abundance (Brown 1984; Gaston *et al.* 2000), the latter is much more difficult to model successfully (see e.g. Nielsen *et al.* 2005), which requires the application of more complex statistical methodologies from those implemented here (Vincent & Haworth 1983; Barry & Welsh 2002; Potts & Elith 2006), and would thus constitute an alternative study approach.

Also, by starting with the premise that steppe bird distributions are determined by factors relating to vegetation type, terrain and disturbance (Osborne *et al.* 2001; Suárez-Seoane *et al.* 2002a), the role of food availability was not considered, even though it has been demonstrated to have an influence on the birds habitat and territory selection (Moreira *et al.* 2004; Traba *et al.* 2008). Nevertheless, the abundance of food (seeds and arthropods) can be associated to specific land use types (Delgado & Moreira 2002) and to the degree of land use intensification (Sousa *et al.* 2004), which have been characterized in the approach used, this way indirectly reflecting food availability. Alternatively, food supply (e.g. arthropod abundance) could be modelled and predicted for being incorporated as a predictor variable in the bird models. Nevertheless, this would imply further field data collection and model building, with the inevitable consequence of error propagation resulting from the combination of the uncertainty of the food supply and bird models.

Moreover, this study analysed the steppe bird community by considering single-species responses to the environment, i.e. by assessing each species individually. Other possible approaches include the assemblage of the species data into the community-level, like e.g. modelling species richness (Cumming 2000; Oindo *et al.* 2003), or the use of multi-response models, capable of analysing multiple species data (Hastie *et al.* 1994; Olden 2003). Indeed the integration of biotic interactions (interactions between species) in the analyses has been seen to improve the prediction of species distributions (Heikkinen *et al.* 2007). On the other hand, care must be taken on the interpretation of such models, as some species might be good predictors of others, due to their co-occurrence and similar habitat preferences instead of true biotic interactions. Also, the use of the occurrence of other species for predicting the target species might improve the model performances, because the former might help explaining environmental factors which were not properly measured before. This means that the occurrence of other species might serve as indirect predictors of the target species, thus making the predictions unreliable (Austin 2002). Still, the implementation of multi-species responses is accessible, as the code used in this work for fitting the MARS models is capable of handling multiple responses, as described by Leathwick *et al.* (2005), although this option was not explored. Ferrier & Guisan (2006) define these two approaches, respectively as "assemble first, predict later" and "assemble and predict together", as opposed to the approach "predict first, assemble later" used on Chapter 2.

The study design – in respect to the definition of the study area location, extent and grain of analysis, at each particular scale – determined the amount of variability contained in the datasets, which is expected to have had an influence on the achieved model performances (Collingham *et al.* 2000; Guisan *et al.* 2007). In particular, the selection of generally suitable areas for the occurrence of the studied species (respectively the Baixo Alentejo region and the Castro Verde landscapes), probably affected the power of discrimination within the models. Indeed, the study by Osborne *et al.* (2001), which analysed Great Bustard occurrence data in the Madrid province (Spain) using a comparable dataset, spatial extent and grain of analysis to that of the present regional study, achieved much higher model performances (AUC score of 0.969). This was probably due to the inclusion of highly unsuitable areas, like urban areas and a mountain range. Also, the work by Suárez-Seoane *et al.* (2002a) on three species in common with the present study, used a similar approach (methodological premises and dataset) and the same grain of analysis (of the regional scale models), while covering a much larger spatial extent (the whole of Spain). This study thus incorporated very high environmental variability which resulted in high model performances (cross-validated AUC of 0.91, on average across the three species). Hence, while restricting the study extent to a narrow environmental range weakens the apparent performance of the models this is due to the fact that the "easy" task of eliminating unsuitable habitats or land cover types has been achieved a priori (Lobo *et al.* 2008).

Osborne & Suárez-Seoane (2002), on the other hand, have noted that partitioning large spatial extents into smaller regions can improve distribution models, by better accounting for regional variations in the dataset. These variations can either refer to region-specific combinations of (indirect) predictors or geographic variation in species behaviour and its interaction with environmental heterogeneity. Another way to deal with these regional variations is to use non-stationary modelling approaches, such as Geographically Weighted Regression (GWR) or Varying Coefficient Modelling (VCM), used in a SDM context by Foody (2004) and Osborne *et al.* (2007). These modelling methods do not assume the modelled associations or processes to be constant across space. These models, however, do not allow for generalisation into other areas (outside the training domain) and do not facilitate the inference of spatially variant species environment associations, with the aim of drawing management recommendations for their conservation.

The selection of the spatial extent of both analyses of the present work, while arbitrarily defined by administrative boundaries (the Baixo Alentejo region and the Castro Verde SPA), nevertheless resulted in areas with particular geographical and environmental characteristics, which were preferred by the studied species. By focusing the analysis within these areas, it was aimed at describing the relevant associations which drive the species occurrences, as well as to identify potential threats to these species, in their stronghold.

In terms of grain of analysis, this study defined it according to the grain of the input species data. While the regional scale data sampling was suited to match the spatial resolution of the remotely sensed imagery used, the predictor variables at the landscape scale were compiled to the area of the species sampling unit (the circular-plot). It has been noted, however, that SDM performance is sensitive to the respective grain of analysis (Guisan *et al.* 2007). Several other studies have made use of geostatistical analysis to assess the grain that best describes the observed patterns of habitat selection (Schaefer & Mayor 2007) and occurrence (Carroll & Pearson 1998). Furthermore, geostatistics can be used for data resampling (by spatial interpolation) to an adequate grain, appropriate for the respective data analysis (Atkinson & Tate 2000). Thus, further work should explore the utility of geostatistical techniques to address this issue.

The presence of high data (multi-)collinearity and concurvity among environmental descriptors, used as predictor variables in multivariate models has been seen to have effects on both model performance and interpretation (MacNally 2002; Morlini 2006; Blanche *et al.* 2008). This study thus followed the general recommendation of removing highly correlated variables before modelling (Freedman *et al.* 1992). In this procedure, however, rank correlations were used, which do not fully capture all non-linear dependencies (concurvities) in the data. Possible methods to deal with data concurvity include the use of non-linear forms of PCAs, such as the Additive PCA (Donnell *et al.* 1994) or the iterative Kernel PCA (Schölkopf *et al.* 1997; Kim *et al.* 2005). These methods, though, would probably increase performance at the cost of interpretability. Nevertheless, their implementation can be extremely useful for the (sole) purpose of prediction.

Alternatively, bootstrapping techniques have also been used to minimize the effects of concurvity in additive models (Figueiras *et al.* 2005). Moreover, when removing inter-correlated variables before modelling, hierarchical partitioning (MacNally 2002; MacNally & Walsh 2004), which is based on a hierarchically exhaustive regression-model building (by running all possible models which include each individual variable), has been used to aid the identification of the subset of variables to retain. As for bootstrapping, this method uses data randomization in order to calculate the significance of the predictors' independent influence on the response variable. During the time of this work, however, none of these methods were possible to implement in a MARS framework. Nevertheless a multi-model approach was used, based on predictor subsets determined by different variable correlation thresholds, this way accounting for some of the (varying) effects of data collinearity in the data fitting.

Data describing environmental systems present high levels of spatial dependencies, or spatial auto-correlation (Legendre 1993; Koenig 1999). As for the non-spatial dependencies (data correlation), spatial autocorrelation can influence the models' performance and interpretation (Segurado *et al.* 2006; Dormann 2007). Common alternative approaches to deal with this are autoregressive models (Augustin *et al.* 1996), geostatistical models (Rossi *et al.* 1992) and GWR (Leung *et al.* 2000). Dormann *et al.* (2007) further explore the issue and reviews it in the context of SDMs. Spatial autocorrelation effects, however, were not accounted for in the present study.

The great potential of the use of remotely sensed data as environmental descriptors in SDMs has been discussed and demonstrated within this study. Pixel-based methods, such as those implemented in the present study, are most commonly used to this end, even though object-based methods, such as image segmentation can be particularly useful for extracting information from high resolution data (Mason *et al.* 2003).

Remote sensing image pixels, however, (due to combined factors like the instrument optics, the detector and electronics) incorporate spectral signal from areas surrounding the nominal instantaneous field-of-view (IFOV). In other words, an image pixel signal includes contributions not only from the field-of-view corresponding to that pixel but also from the neighbouring areas, in a way described by a point spread function (PSF), which is sensor-dependent (Townshend 1981; Cracknell 1998; Huang *et al.* 2002b). If the respective sensor's PSF is known, it can be inverted to estimate the true ground response by image enhancement – de-convolution of the observed response (Forster & Best 1994). In most cases, however, this is not feasible and this remains a source of uncertainty in the data, which has been noted by Townshend *et al.* (2000) to influence land-use characterisation. These authors thus recommend that land cover properties be reported at spatial resolutions coarser than the individual pixel. Additionally, interpolation procedures such as image resampling applied for geometrical rectification, constitute an additional source of uncertainty resulting from image processing methodologies (Cracknell 1998). Indeed, image registration errors have been widely associated with biases in change detection studies (e.g. Townshend *et al.* 1992; Dai & Khorram 1998; Wang & Ellis 2005). In the context of the use of remote sensing imagery in SDMs, Osborne & Leitão (*in press*) (see Appendix A.2) have seen these type of errors to impact on the models' performances, predictions and interpretations. These authors further recommend that researchers, as a rule of thumb, restrict analysis to grain sizes at least twice that of the largest likely error in the datasets. To this concern, the grain sizes used in the analyses in the present work (and particularly so in the landscape scale models), were much greater than the expected errors in the original input data.

The hierarchical bottom-up MS model integration method used, i.e. the incorporation of the predictions from the coarser scale as descriptive variables in the finer scale models is not new (see e.g. Pearson *et al.* 2004). It is, however, one of many different possible approaches to MS data analysis. For example, another similar top-down approach consists of using coarse-scale species occurrence (or atlas) data instead of model predictions (Araújo *et al.* 2005; McPherson *et al.* 2006). Alternatively, top-down approaches can make use, for example, of aggregated fine-scale species response to heterogeneity for explaining the observed coarse scale patterns (With & Crist 1996). Yet another approach is to incorporate predictor variables collected at different scales (grain sizes) in a single model (Estes *et al.* 2008). In such cases, the decomposition of scale effects can be assessed with the use of methods like hierarchical variance decomposition (Cushman & McGarigal 2002). There is, however, no reason to assume that some approaches are any better than others for describing patterns across scales.

6.3. *Future lines of research*

Finally, several topics (besides those covered in this discussion) arose during the course of this work, which need further investigation. Some of them, either by their relevance or by the set-up assembled in this work (e.g. datasets, tools, knowledge), would be most suited as follow-up research.

For example, the collection of high-resolution multispectral (CASI) data during the STEPPEBIRD flight, together with the LiDAR data and the extensive bird datasets collected, allows the exploration of the use of high-resolution RS data for describing landscape heterogeneity in these habitats.

Also, during the spring of 2005, during the most severe drought of the last 60 years in Portugal, a bird dataset was compiled (990 data points), which is equivalent to the one used in the landscape scale models presented. These two datasets (2005 and 2006), together with the respective Landsat satellite imagery (also acquired), make up ideal datasets for answering ecological questions about the steppe birds in the region, which could potentially be extrapolated for other taxa and other regions. One such question could be: are the species habitat associations explained by the SDMs transferable to different environmental conditions, within the same area? Or: do the patterns of species co-occurrence remain constant during severe climatological events? Both the previous questions have clear implications for climate change research. Unfortunately, an attempted application for a NERC small grant presenting these questions in this context (with the aim of funding the analysis of these data) was rejected with the claim that single drought events are not related with climate change issues, which is highly arguable. With this claim, however, the application assessment report did not consider the unique opportunity of the existence of such an ideal real-life dataset for addressing those questions, even though it did not dispute the quality and emergence of the scientific topic presented.

Additionally, and deriving from the previous proposed question, the development of community models (such as multi-response models) could answer further questions about the structure of the steppe bird community in the region, as well as identify relevant biotic interactions between the studied species.

As discussed in Chapter 5.4, the apparent effect of species body size on the grain of the fitted responses for each species raises macroecological questions which would be interesting to expand. Indeed, the study by Suárez-Seoane *et al.* (2002a) focussed on three steppe bird species of highly contracting body sizes (Great Bustard, Little Bustard and Calandra Lark), but by fixing the grain of all environmental predictors (at 1 x 1 km^2 pixel resolution) was unable to address this question.

The integration of these datasets with those existing for the rest of the country and for Spain (and possibly also for Morocco), eventually with the collection of some more field data (in areas with low data density) would allow the exploration of biogeographical questions. For example, while only covering Spain, Suárez-Seoane *et al.* (2002b) found that birds of different biogeographical origins showed different responses to agricultural land abandonment, which has clear consequences for species conservation.

More importantly, the translation of the model results into conservation practices would be desirable. This could be done either through the drafting of general management recommendations based on the habitat associations described, or by using the species distribution predictions for identifying potentially important steppe bird areas which are currently unprotected. The latter could, for example, make use of reserve design algorithms such as those used by Cabeza *et al.* (2004). Particularly, it would be interesting to understand if EU policies targeting species conservation is perhaps failing because they are acting at the wrong scale. Indeed, Whittingham (2007) suggests that EU Agri-environmental schemes are applied to very small patches of land, this way creating a complex mosaic of differing habitat quality at a larger scale.

Lastly, the incorporation of spatial (landscape) context variables relating to habitat configuration, fragmentation or connectivity (Opdam 1991; McGarigal & McComb 1995), could also further help explain the observed patterns of species distributions, as well as give more insights into species population dynamics, for example by associating particular landscape structural features to decreasing population densities (García *et al.* 2007). Further research in this area could improve the understanding of the factors influencing these species, also with clear implications for their conservation (Hansson & Angelstam 1991).

APPENDICES

A.1. Spatial distribution patterns, habitat correlates and population estimates of steppe birds in Castro Verde

A.1.1. Abstract

Castro Verde is the main area of cereal steppes in Portugal (ca. 80,000 ha), having international importance for several steppe bird species with unfavourable conservation status. In spring 2006, a large-scale assessment of bird populations in the region was carried out using a simple methodological procedure. The occurrence and abundance of 16 species of steppe birds was estimated in 391 squares (1 x 1 km) in order to describe the spatial distribution patterns, explore the habitat variables explaining the observed patterns and estimate population sizes. The more frequent steppe species in the region were Corn Bunting *Miliaria calandra* (present in 78% of the sampling points), Calandra Lark *Melanocorypha calandra* (29%), Crested / Thekla larks *Galerida* spp. (29%) and Little Bustard *Tetrax tetrax* (28%). In terms of estimated population sizes, we confirmed the national importance of Castro Verde for several species, most noticeably Great Bustard *Otis tarda*, Little Bustard, Calandra Lark and Montagu's Harrier *Circus pygargus*. Regarding species habitats associations, four groups of species could be identified: a) those associated with fallow land and grasslands, e.g. Calandra lark; b) species associated with cereal fields, e.g. Zitting Cisticola *Cisticola juncidis*; c) species associated with ploughed fields, e.g. Black-eared Wheatear *Oenanthe hispanica*; and d) species associated with habitat mosaics, e.g. Galerida larks. Although simple, the methodology used permitted the characterization of the present distribution and abundance patterns, and established a baseline for the monitoring of changes in the future.

A.1.2. Introduction

The pseudosteppes of the Iberian Peninsula are one of the farmland habitat types holding a larger number of bird species with unfavourable conservation status (Suárez *et al.* 1997). Pseudosteppes occupy an area over 4,5 million hectares (Suárez *et al.* 1997) representing an important part of the Natura 2000 network in the region. There are several types of pseudosteppes, including semideserts, páramos and cereal steppes (Tellería *et al.* 1988; Martinez & Purroy 1993), but the latter is more common in western Spain and Portugal, and holds populations of many endangered birds including globally threatened species such as the Great Bustard (*Otis tarda*) and the Lesser Kestrel (*Falco naumanni*) (Tucker & Heath 1994; Tucker 1997).

Cereal steppes result mostly from the cultivation of dry cereal crops and extensive pastures. Thus, they are economically marginal farming systems threatened by agricultural intensification in the more productive soils, agricultural abandonment, often with afforestation of agricultural land, in poorer soils and, more generally, changes in management practices according to agricultural policy trends (Suárez *et al.* 1997).

Castro Verde is the main area of cereal steppes in Portugal (Costa *et al.* 2003). It has national and international importance for populations of several steppe bird species including Great Bustard, Little Bustard (*Tetrax tetrax*), Calandra Lark (*Melanocorypha calandra*), Lesser Kestrel, Stone Curlew (*Burhinus oedicnemus*), Roller (*Coracias garrulus*) and Black-bellied Sandgrouse (*Pterocles orientalis*) (Costa *et al.* 2003). It is the most important area in the country for Great Bustard and Lesser Kestrel (Costa *et al.* 2003; Pinto *et al.* 2005) and it holds high densities of breeding Little Bustards (the highest in Europe) and Calandra Lark (Moreira 1999). As other steppe regions, Castro Verde is threatened by changes in farming practices and agricultural abandonment. Owing to its ornithological importance, three LIFE projects on steppe bird conservation have been carried out in the region. Moreover, a specific agri-environmental programme for farmers in the area (Castro Verde Zonal Plan) allows the existence of subsidies to carry out agricultural practices compatible with bird conservation.

Scientific research in Castro Verde was scarce until the 1990's, although the monitoring of some species such as the Great Bustard (Pinto *et al.* 2005), Lesser Kestrel (Rocha *et al.* 1996) and Crane *Grus grus* (Almeida 1992) had started since the 1980's. The beginning of the first LIFE project carried out by the Portuguese League for Nature Conservation (LIFE92 NAT/P/013900 – First phase of the conservation of steppic birds in Castro Verde) boosted scientific research in the area in the late 1990's, either on bird communities (Leitão & Moreira 1996; Moreira & Leitão 1996a, b) or single-species studies (Franco *et al.* 1996; Morgado & Moreira 2000). Since then, there have been a growing number of scientific publications on Castro Verde's birds. In spite of this wealth of information there has been no large-scale detailed assessment of distribution patterns, or population estimates, for most species in the region. This is a drawback that hinders an effective characterization of the current situation (or system state) and the monitoring of likely changes in the near future (Yoccoz *et al.* 2001; Martin *et al.* 2007).

As part of a EUFAR (European Fleet for Airborne Research) research project (STEPPEBIRD), the Natural Environment Research Council (UK) NERC undertook flights over Castro Verde in spring 2006 to collect detailed remotely-sensed data on habitat type and vegetation structure. These data are still being processed and will be used in further scientific studies, including bird-habitat relationships and vegetation structure. The opportunity was also taken to carry out a large-scale census of bird populations in the region using point counts to: a) describe the spatial distribution patterns of ground nesting steppe birds in the region; b) explore the habitat variables explaining the observed patterns; c) obtain population estimates for the more common species. The final aim was to provide a baseline characterisation against which the results of future bird monitoring, using a similar approach, could be compared.

As a final remark, one must notice that this study was carried out in the spring of 2006, the year following the worst drought of the last 60 years in Portugal. In fact, during 2005 rainfall in the region was just 40% of an average year (INAG 2005), which resulted in a poor agricultural year, particularly for dry crops (cereal yield was very low). This drought probably had important impacts on bird populations, mainly for resident species, which are likely reflected in the current results. This should be born in mind when discussing the present results and comparing them with future surveys.

A.1.3. Methods

- *Study area*

The Castro Verde special protection area (SPA; Figure A.1.1) is a plain (100–300 m) of about 80,000 ha, having a Mediterranean climate including hot summers (30–35°C on average in July), fairly cold winters (averaging 5-8°C in January) and over 75% of annual rainfall (500–600 mm) concentrated in October–March (Delgado & Moreira 2000; Moreira *et al.* 2005). It is mainly occupied with pseudo-steppe habitats (ca. 55,000 ha; Figure A.1.2) created by farming activities. The traditional agricultural system used in this region is as follows: each farm is divided into parcels, each lying under cereal cultivation for two years, after which the land is left fallow, normally for 2–3 years. The parcel is then ploughed to re-initiate the rotation cycle. Fallow land is generally used as pasture for sheep and, more rarely, cattle. In the north and south of the region there are holm oak *Quercus rotundifolia* woodlands ('montados') of scarce tree cover, frequently with a grassy understory grazed by livestock. Other forested areas are more rare and include olive groves, old eucalyptus *Eucalyptus spp.* plantations and recent (<10 years) afforestations with eucalyptus, umbrella pines *Pinus pinea* and holm and cork oak *Quercus suber* (Figure A.1.2).

Figure A.1.1 - Location of the Castro Verde Special Protection Area for birds, and location of main roads, rivers and villages

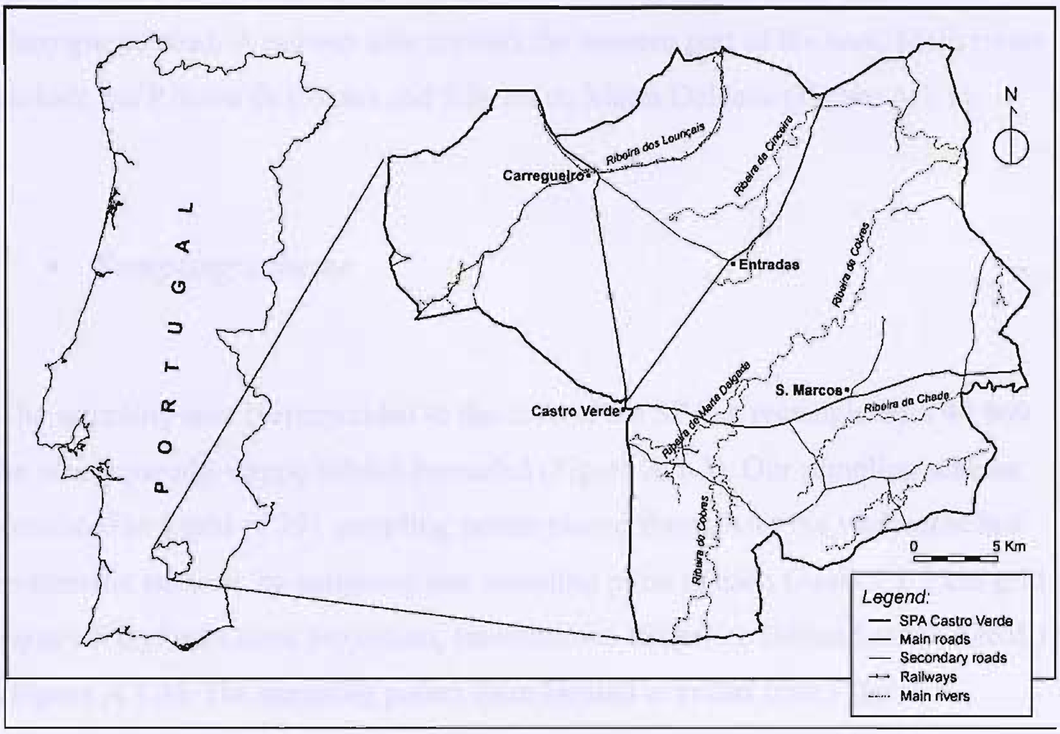
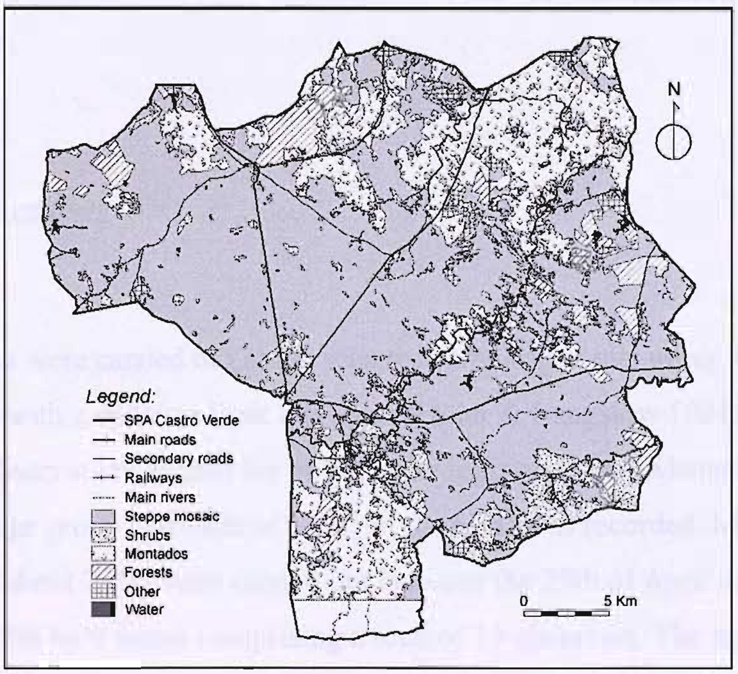


Figure A.1.2 - Land use map of the Castro Verde Special Protection Area. Adapted from Project LIFE2002/NAT/P8481 "Recuperação do Peneireiro-das-torres (Falco naumanni) em Portugal"



Areas of shrublands occur mainly in association with river valleys and in the south-eastern part of the region, as a mosaic of shrublands interspersed with old fallows resulting from agricultural abandonment and scrub encroachment

(Moreira *et al.* 2005). Three main roads cross the area, the Castro Verde – São Marcos road, the Castro Verde – Entradas road, and the Castro Verde – Carregueiro road. A railway also crosses the western part of the area. Main rivers include the Ribeira de Cobres and Ribeira de Maria Delgada (Figure A.1.1).

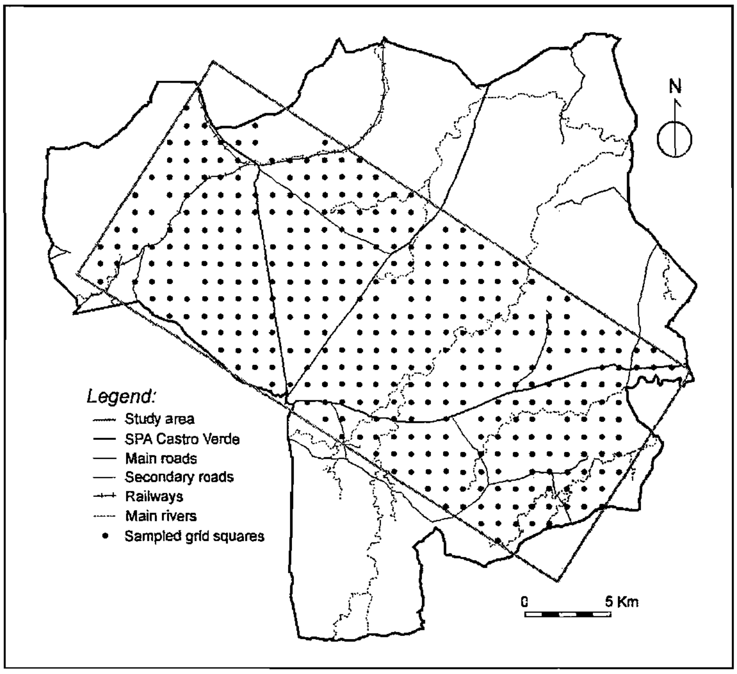
- *Sampling scheme*

The sampling area corresponded to the core of the SPA, a rectangle with 44,860 ha where pseudo-steppe habitat prevailed (Figure A.1.3). Our sampling scheme consisted in a grid of 391 sampling points placed throughout the study area in a systematic manner, by assigning one sampling point to each *Gauss* 1 x 1 km grid square (Hayford-Gauss projection, International Ellipsoid, Datum Lisboa IGeoE) (Figure A.1.3). The sampling points were located over dirt tracks (for accessibility) and as close as possible to the square's centre. A 125m circular buffer was defined around each point, and it was also required that this buffer fell completely on pseudo-steppe habitat (based on Figure A.1.2) and within a single grid square. In cases where these conditions did not apply, the grid square was not surveyed.

- *Bird counts*

Bird censuses were carried out at the selected sampling points using 5-minute point counts with a distance limit of 125 m (Fuller & Langslow 1984; Bibby *et al.* 2000). All observations within the buffer were registered and, whenever possible, the sex and age group (juvenile or adult) of the birds was recorded. Most of the bird counts (about 75%) were carried out between the 29th of April and the 8th of May of 2006 by 9 teams comprising a total of 19 observers. The remaining counts were carried out in a larger time span (between the 20th of March and the 12th of May) by two observers (PJL and RM) of the former group. All counts were carried out in the first 4 hours after sunrise and in the last 2 hours before sunset. Categorisation to the genus level was made for the Crested and Thekla larks (*Galerida cristata* and *Galerida theklae*) due to difficulties in reliably identifying all individuals of these two species in the field. All observers were experienced, thus we believe inter-observer differences did not significantly affect the results. A joint session with all observers to improve the team accuracy in estimating the distance limit for bird counts was carried out prior to the surveys.

Figure A.1.3 - Study area (rectangle) and location of survey points for bird counts and habitat measurements



- *Habitat measurements*

Land-use information was collected in each sampling point by dividing the 125m-radius buffer into 8 quadrants and visually estimating the dominant habitat in each one of them. The following seven habitat categories were considered: a) fallow land and grasslands; b) fallow land and grasslands with scattered shrubs; c) cereal fields; d) ploughed fields; e) shrublands; f) afforestations; g) holm oak 'montados'.

- *Data analysis*

In total, ca. 3000 birds from 62 different species were observed. For the purposes of this study, we selected only 16 species (Table A.1.1) including mostly ground-nesters, but also non-obligate ground nesters that, although not exclusive of steppe-like habitats, were fairly common in the study area. The abundance of each species in each point was expressed as: (a) the number of pairs for songbirds (Passeriformes) and Quail (*Coturnix coturnix*). The number of pairs was determined using only the number of singing males, unless twice that number was less than the number of singing males plus the number of all other observations. In the latter case, the number of pairs was determined from half the sum of the number of singing males plus the number of all other observations (DeSante 1981); (b) number of males for Little Bustard (*Tetrax tetrax*). In this species with a polygynous mating systems, females are quite inconspicuous and male density is usually assessed for population monitoring (Faria & Rabaça 2004); and (c) as number of individuals for the remaining species, in which difficulties in separating males from females occurred, or total population is usually assessed without discriminating gender. Population estimates were derived simply by using the mean and 95% confidence intervals of bird density in each point, for each species, to extrapolate to the total steppe area in the region (55,490 ha). Without correction for detectability, the obtained values are probably underestimates, and should be used only to compare time variations within-species; between-species comparisons should not be made. The average population size was compared with the estimates given by BirdLife International (BirdLife International 2004) for Portugal, in order to assess the national importance of Castro Verde.

For each species, we produced a map showing all the points where the species occurred (presence/abundance). For visualisation purposes and interpretation of the spatial pattern of the four most frequent species, we interpolated the presence data points using Ordinary Kriging, a geostatistical technique capable of producing probability contour maps, derived from point data (Rossi *et al.* 1992; Burroughs 1995). For the more common species (Corn Bunting), an interpolated map of abundance (pairs/point) was also produced, using the same technique. All the geostatistical analyses were carried out in ArcGIS version 9.0 software package (ESRI 2004).

Table A.1.1 - List of the 16 species studied, ordered by decreasing frequency of occurrence in the 391 sampled points

Scientific name	Common name	Proportion of points
Miliaria calandra	Corn Bunting	0.783
Melanocorypha calandra	Calandra Lark	0.294
Galerida spp.	*Galerida* larks	0.289
Tetrax tetrax	Little Bustard	0.276
Saxicola torquata	Stonechat	0.148
Cisticola juncidis	Zitting Cisticola	0.113
Alectoris rufa	Red-legged Partridge	0.100
Calandrella brachydactyla	Short-toed Lark	0.097
Circus pygargus	Montagu's Harrier	0.066
Coturnix coturnix	Quail	0.066
Burhinus oedicnemus	Stone Curlew	0.064
Upupa epops	Hoopoe	0.061
Anthus campestris	Tawny Pipit	0.054
Oenanthe hispanica	Black-eared Wheatear	0.051
Otis tarda	Great Bustard	0.041
Pterocles orientalis	Black-bellied Sandgrouse	0.020

From the seven initial habitat variables describing habitat availability (number of quadrants where the habitat was dominant, ranging from zero to eight), other variables were derived: presence of habitat (binary variable stating if the habitat was present in any of the eight quadrants), habitat dominance (binary variable stating if the habitat was dominant on four-or-more quadrants), and habitat richness (total number of habitat classes in the buffer). Habitat dominance variables were only extracted in cases where, through empirical inspection of the bird and habitat data (by using scatter plots), we found some significant pattern that could explain the bird's probability of occurrence.

In order to explain the species-habitat associations, we applied a Generalized Linear Model (GLM) with a logit link function (logistic regression) using species presence-absence data (derived from the field data) and the habitat variables described above. In order to reduce the number of variables to enter in the models, we first used a Mann-Whitney U test (univariate non-parametric test for independent samples) to screen the variables excluding those that showed a weak association with the bird presence absence data ($p > 0.1$). In the multivariate GLMs, we used a forward stepwise (likelihood ratio) variable selection method as an exploratory approach. Based on the chosen model (by the stepwise procedure), we included and excluded variables considered important to describe the bird's probability of occurrence, based on ecological knowledge of the species, and compared the several models obtained. We used an Information Theoretic (IT) approach based on the AIC values to choose the best model for each species (Akaike 1974; Burnham & Anderson 2002, 2004).

A.1.4. Results

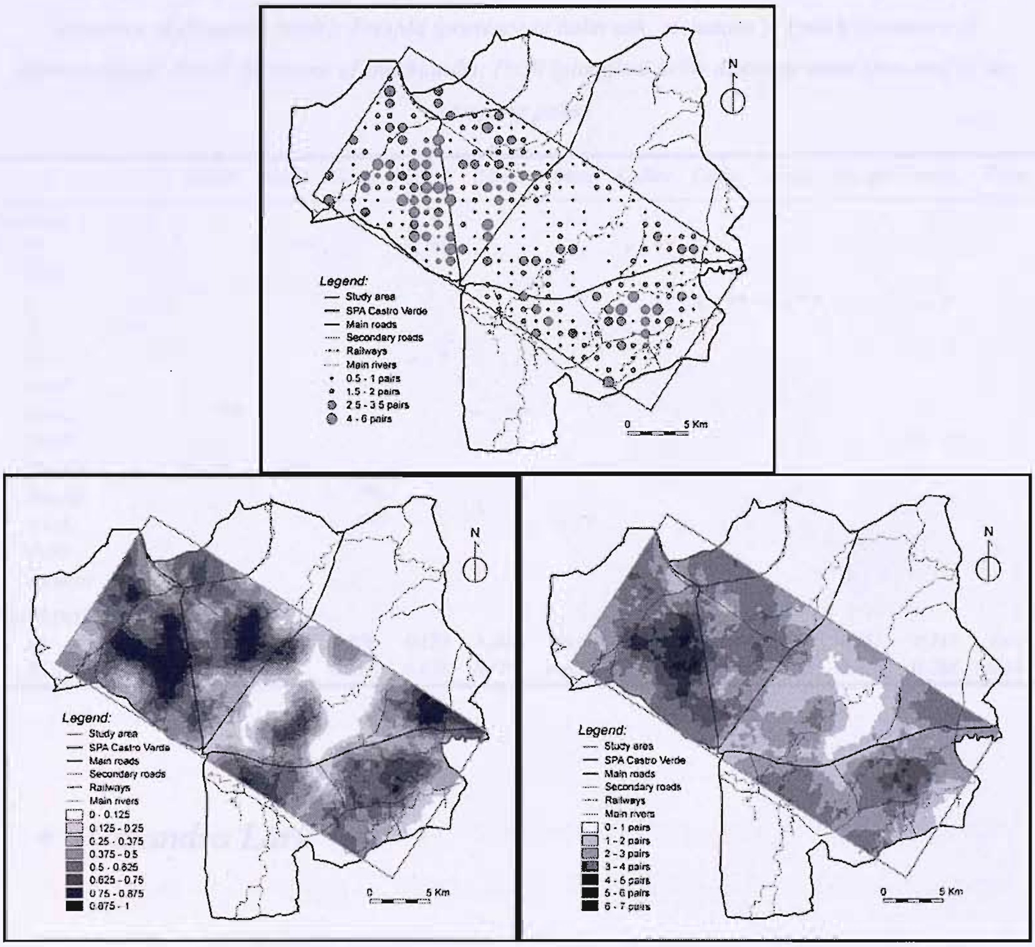
The frequency of occurrence of the species studied is shown in Table A.1.1. Results are detailed below, ordered by decreasing species frequency.

- *Corn Bunting*

The Corn Bunting was by far the most frequent species in the study area, occurring in 78.3% of the sampling points, with an average abundance of 1425 pairs per point (range = 0-6; SE = 0.060), which yielded an estimate of 16185 (95% CI = 14852-17519) pairs for the total pseudo-steppe area of the Castro Verde SPA. The national population of this species has been roughly estimated as 100,000 to 1,000,000 pairs (BirdLife International 2004), so it is difficult to evaluate the significance of the Castro Verde population, as it could range from 2 to 16% of the national population.

The species was very common and widespread in the region, although more likely to be found west of the Castro Verde - Entradas road. Two hotspots with higher population densities were identified: west of the Castro Verde – Carregueiro road and in the southeast, south of Ribeira da Chada (Figure A.1.4).

*Figure A.1.4 - Map of Corn Bunting (*Miliaria calandra*) occurrence (above), interpolated map of probability of occurrence (below left) and interpolated map of abundance (below right). Sample points where the species did not occur are not shown, and dot size is proportional to abundance. The darker the colour the higher the probability of occurrence/estimated abundance*



The existence of cereal fields was the main factor increasing Corn Bunting probability of occurrence (Table A.1.2). The species was also more likely to be found near 'montados'. The association of the species with cereal fields during the breeding season has already been reported (Delgado & Moreira 2000; Stoate *et al.* 2000) and is probably related to the availability of food (mainly arthropods) combined with appropriate nest cover provided by tall and dense vegetation. The presence of scattered holm/cork oak trees in the 'montados' probably increases breeding habitat suitability, as these are intensively used as perches by singing males. Furthermore, many 'montados' have an understory of cereal crops.

Table A.1.2 - Summary of logistic regression models, indicating for each species, the habitat variables selected, their slope (positive or negative), and significance values (p<0.05; ** p<0.01; and *** p<0.001). Model performance is indicated by the Nagerkelke r2 and the area under the ROC curve. Variable legend: F (amount of fallow land and grasslands); F+S (amount of fallow land and grasslands with scattered shrubs); C (amount of cereal fields); P (amount of ploughed fields); Md (amount of holm oak 'montados'); PresF (presence of fallow land and grasslands); PresC (presence of cereal fields); PresP (presence of ploughed fields); PresMd (presence of holm oak 'montados'); PresAf (presence of afforestations); PresF (presence of shrublands); Po50 (ploughed fields dominate more than half of the sampling point)*

	Milcal	*Melcal*	*Galsp*	*Tettet*	*Cisjun*	*Aleruf*	*Calbra*	*Cirpyg*	*Cotcot*	*Upaepo*	*Oenhis*	*Pteori*
Variables												
F		+ , ***		+ , ***			+ , **					
F+S								+ , **				
C								+ , **	+ , ***		– , *	
P							+ , **					
Md				– , *								
PresF			+ , ***									
PresC	+ , ***				+ , ***							
PresP											+ , ***	
PresMd	+ , *	– , **				+ , **						
PresAf			+ , **									
PresS			+ , ***			+ , *						
Po50												+ , *
Richness										+ , *		
Model performance												
r2	*0.109*	*0.208*	*0.179*	*0.122*	*0.200*	*0.059*	*0.108*	*0.080*	*0.114*	*0.031*	*0.147*	*0.043*
ROC AUC	*0.669*	*0.731*	*0.696*	*0.676*	*0.743*	*0.628*	*0.671*	*0.694*	*0.728*	*0.624*	*0.701*	*0.557*

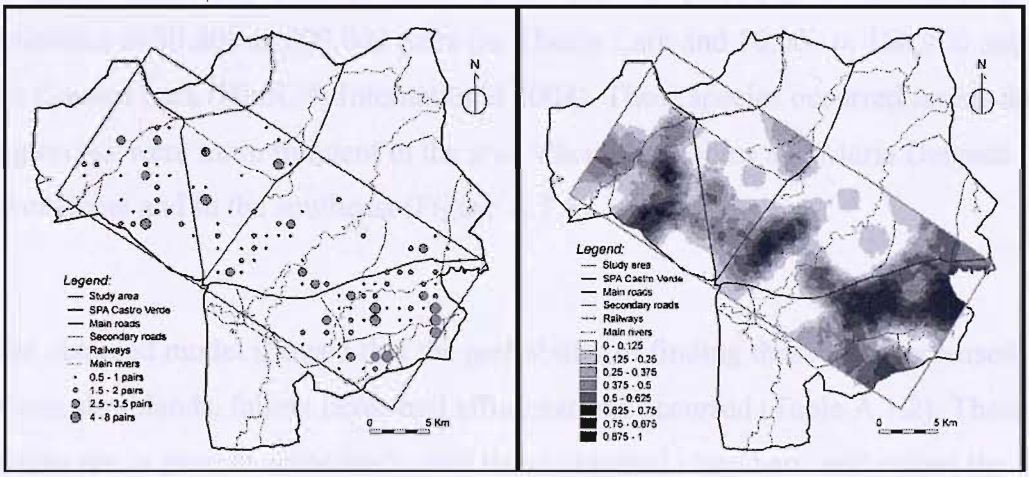
- *Calandra Lark*

Calandra Lark was the second most frequent species, being present in 29.4% of the sampled points, with an average abundance of 0.542 pairs per point (range = 0-8; SE = 0.056), which yielded an estimate of 6160 (95% CI = 4910-7410) pairs for the total pseudo-steppe area of the Castro Verde SPA. When compared with the previous population estimate for this region (400-8500 pairs, (Costa *et al.* 2003), this new estimate adds extra precision, particularly in the lower range limit, and according to national population estimates (BirdLife International

2004), represents at least 60% of the total Portuguese population, showing how important this area is for the conservation of the species.

The Calandra Lark occurred all over the region, although three main nuclei of occurrence could be identified. The largest one was in the southeast (south of the Castro Verde-São Marcos road). A second was found between the Castro Verde-Entradas road and the Cobres River, and the third was located west of the Castro Verde - Carregueiro road (Figure A.1.5).

*Figure A.1.5 - Map of Calandra Lark (*Melanocorypha calandra*) occurrence (left) and interpolated map of probability of occurrence (right). Sample points where the species did not occur are not shown, and dot size is proportional to abundance. The darker the colour the higher the probability of occurrence/estimated abundance*



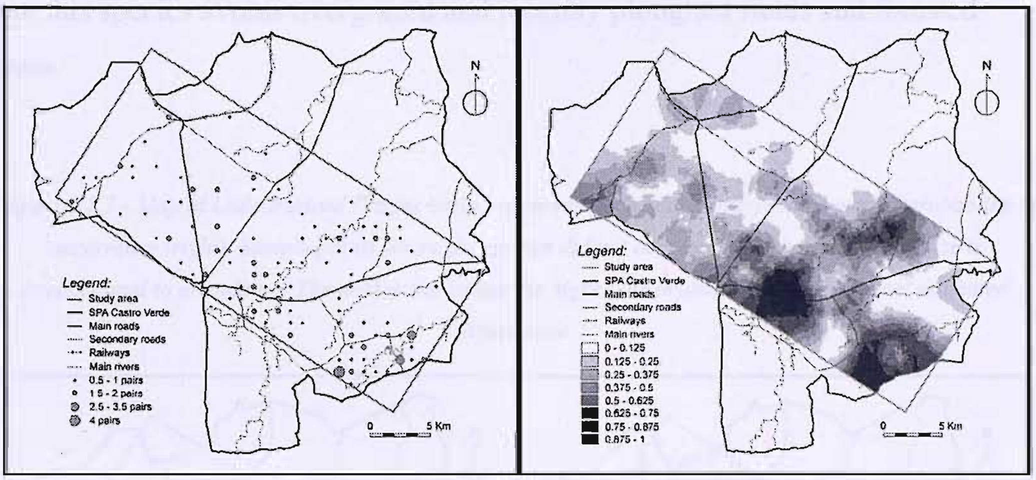
Calandra Larks were more likely to be seen in points having a higher proportion of fallow fields and in areas without 'montados' (Table A.1.2). The association of the species with fallow fields during the breeding season in this region is in agreement with the results of previous studies (Moreira & Leitão 1996b; Moreira 1999; Delgado & Moreira 2000), and could be related to the characteristic vegetation structure and diversity of fallow fields, which seems to suit better the breeding ecological requirements this lark species (e.g. nest cover, food availability and accessibility, predation risk perceiving) in comparison to alternative breeding habitats (e.g. cereal fields, ploughed fields). The species' avoidance of forested areas, including 'montados', had already been recognized (e.g. Cramp 1988).

- *Crested / Thekla Larks*

Crested / Thekla Larks were present in 28.9% of the sampling points, with an average abundance of 0.329 pairs per point (range = 0-4; SE = 0.031), which yielded an estimate of 3734 (95% CI = 3031-4437) pairs for the total pseudo-steppe area of the Castro Verde SPA. Although these two species were not separated in the field due to their morphological similarities, most observations should correspond to Thekla Larks as they are more abundant in eastern Alentejo than Crested Larks (Rufino 1989). The Portuguese population has been roughly estimated as 50,000 to 500,000 pairs for Thekla Lark and 10,000 to 100,000 pairs for Crested Lark (BirdLife International 2004). These species occurred across the region but were more frequent in the area where the Cobres and Maria Delgada rivers meet and in the southeast (Figure A.1.6).

The obtained model showed that the probability of finding these larks increased where shrublands, fallow fields and afforestations occurred (Table A.1.2). These results are in general accordance with those obtained elsewhere, and reflect the association of these species to heterogeneous environments and to the presence of shrub-like cover. Manrique & Yanes (Manrique & Yanes 1994) for example, describe optimum habitat for Thekla Larks as open scrub of low to medium height in arid or semi-arid terrain. Rufino (Rufino 1989) reports that in agricultural habitats, Thekla Larks occupy fallow land with scattered shrubs and trees, and sparse holm oak 'montados'. In Castro Verde, Delgado & Moreira (Delgado & Moreira 2000) did not find a clear association of Galerida larks with specific habitat types, but other studies on fallow land showed that they prefer grasslands with scattered shrubs or trees (Moreira 1999; Santos 2000; Moreira *et al.* 2005).

*Figure A.1.6 - Map of Crested/Thekla Larks (*Galerida *spp.) occurrence (left) and interpolated map of probability of occurrence (right). Sample points where the species did not occur are not shown, and dot size is proportional to abundance. The darker the colour the higher the probability of occurrence/estimated abundance*



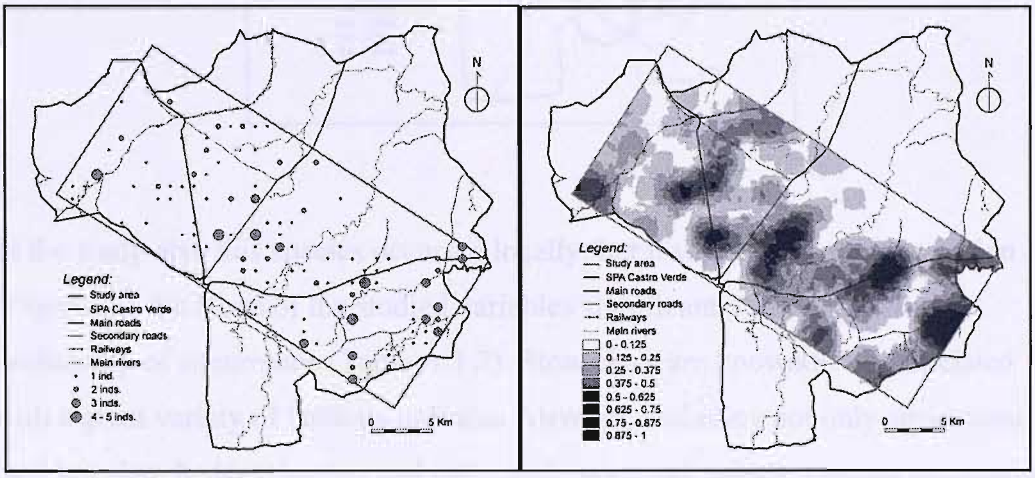
- *Little Bustard*

The Little Bustard occurred in 27.6% of the sampling points, with an average abundance of 0.371 males per point (range = 0-4; SE = 0.036), which yielded an estimate of 4213 (95% CI = 3402-5025) male Little Bustards for the total pseudo-steppe area of the Castro Verde SPA. These figures represent roughly 24% of the most recent estimate for the Alentejo region (17551 displaying males; Silva *et al.* 2006). Considering that the Alentejo holds 85% of the distribution area of Little Bustard in Portugal (Rufino 1989; Silva *et al.* 2006), we can say that the study area supports around 20% of the national population of this species. The estimate presented here is higher than the most recent estimate for Castro Verde (3340 displaying males, (Silva *et al.* 2006).

The species occurred in the whole study area, although four to five scattered nuclei with higher prevalence could be identified (Figure A.1.7). The main factor influencing (positively) the probability of occurrence was the availability of fallow fields. In contrast, a higher availability of 'montados' decreased the probability of Little Bustard occurrence (Table A.1.2). This agrees with the patterns found in other studies conducted in Castro Verde and in other areas of

Alentejo (Moreira & Leitão 1996a, b; Moreira 1999; Faria & Rabaça 2004). These studies clearly showed that the Little Bustard is strongly associated with the grass layer of large fallow fields, both for displaying and laying the eggs, and that this species avoids overgrazed and recently ploughed fields and forested areas.

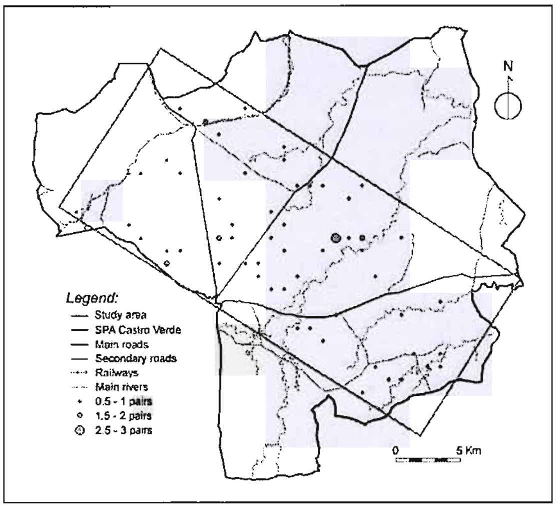
*Figure A.1.7 - Map of Little Bustard (*Tetrax tetrax*) occurrence (left) and interpolated map of probability of occurrence (right). Sample points where the species did not occur are not shown, and dot size is proportional to abundance. The darker the colour the higher the probability of occurrence/ estimated abundance*



- *Stonechat*

The Stonechat occurred in 14.8% of the sampling points, with an average abundance of 0.136 pairs per point (range = 0-3; SE = 0.019), which yielded an estimate of 1540 (95% CI = 1122-1958) pairs for the total pseudo-steppe area of the Castro Verde SPA. Recent studies carried out in the area did not provide information on either abundance or population estimates (Moreira & Leitão 1996a, b; Delgado & Moreira 2000), thus the current estimates are the first available for the region. Rufino (Rufino 1989) suggested that this species is more abundant in the Alentejo region and Beira Interior than in the rest of the country. In any case, the Castro Verde population is not a significant proportion (at most ca. 5%) of the national population (25,000-250,000 pairs; BirdLife International 2004).

*Figure A.1.8 - Map of Stonechat (*Saxicola torquatus*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

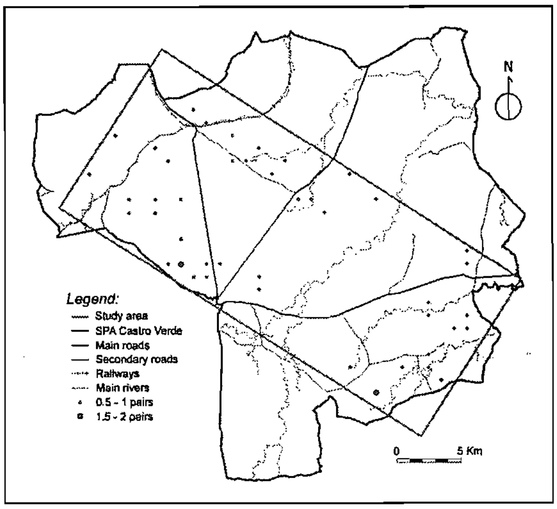


In the study area this species occurred locally, but it was scattered in the region (Figure A.1.8). None of the studied variables significantly influenced its probability of occurrence (Table A.1.2). Stonechats are known to be associated with a great variety of habitats in Baixo Alentejo, including not only agricultural land but also, hedges, bushes and salt marsh, open oak woods, riverine vegetation and even sand dunes with bushes (Soares 1999). It is likely that other local and landscape variables that were not measured in this study, such as the amount of edges, fragmentation variables or vegetation structure, influence the Stonechat's distribution.

- *Zitting Cisticola*

The Zitting Cisticola was present in 11.3% of the sampling points, with an average abundance of 0.115 pairs per point (range = 0-2; SE = 0.017), which yielded an estimate of 1308 (95% CI = 931-1684) pairs for the total pseudo-steppe area of the Castro Verde SPA. In comparison with the estimate for the national population (50,000-500,000 pairs; BirdLife International 2004), the population in Castro Verde is not significant at the national level.

*Figure A.1.9 - Map of Zitting Cisticola (*Cisticola juncidis*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



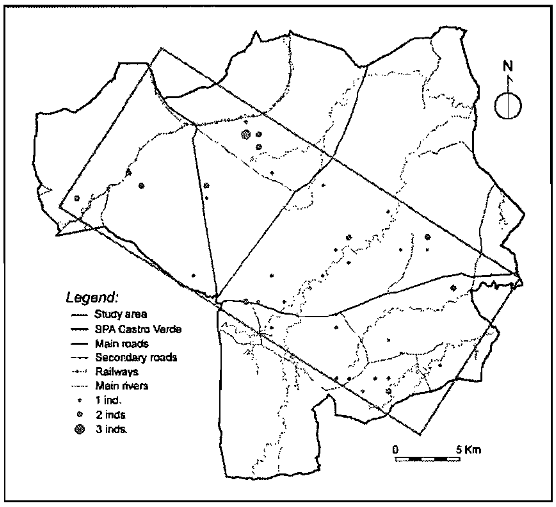
This species could be found scattered in the area (Figure A.1.9). It was more likely to occur in points where cereal fields were present (Table A.1.2), which agrees with previous studies that have showed that the species is much more abundant in cereal fields than in other habitat types (Delgado & Moreira 2000). Delgado & Moreira (Delgado & Moreira 2002) also found a preference of Zitting Cisticola for wheat over barley and oat fields, suggesting that the incorporation in this analysis of variables describing cereal type structure would produce a more accurate prediction model of this species' occurrence.

- *Red-legged Partridge*

The Red-legged Partridge was detected in 10.0% of the sampling points, with an average abundance of 0.133 birds per point (range = 0-3; SE = 0.022), which yielded an estimate of 1511 (95% CI = 1022- 2000) individuals for the total pseudo-steppe area of the Castro Verde SPA. This is equivalent to a density of ca. 0.02 partridges/ha, similar to the estimates of Borralho *et al.* (1997; Borralho *et al.* 2000) in other areas without specific game management in Alentejo region. The Portuguese Red-legged Partridge population was estimated in 10,000-

100,000 pairs (BirdLife International 2004), indicating a low importance of Castro Verde SPA population in a national context (less than 5%).

*Figure A.1.10 - Map of Red-legged Partridge (*Alectoris rufa*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

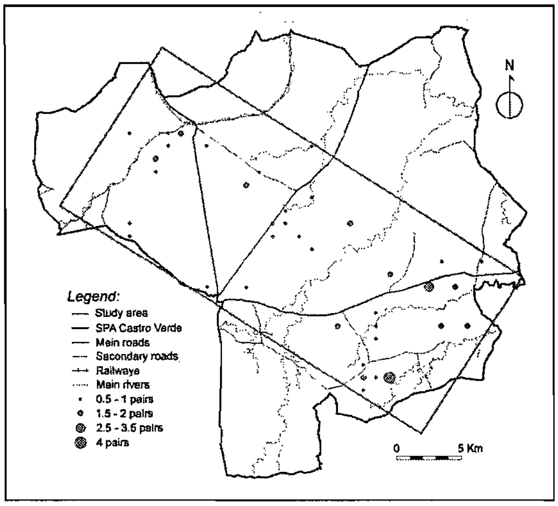


The species has a scattered distribution in the region (Figure A.1.10), being more common where holm / cork oak trees ('montados') and shrublands were present (Table A.1.2). The positive association with scattered 'montados' and shrubs may be related with the availability of shelter and breeding sites. This has been already reported in similar areas, with partridges showing a preference for boundaries and shrub patches (Fortuna 2002). Although Borralho *et al.* (1999) found a positive association with fallows during the breeding season, in this study this association was not found, probably due to the sampling period (late for this species).

- *Short-toed Lark*

The Short-toed Lark occurred in 9.7% of the sampling points, with an average abundance of 0.128 birds per point (range = 0-4; SE =0.023), which yielded an estimate of 1453 (95% CI = 939-1966) pairs for the total pseudo-steppe area of the Castro Verde SPA. If compared with the population estimates of 2000-20,000 pairs for Portugal (BirdLife International 2004), this corresponds at least to 7% of the national population, but could be as high as 60% or more. There are no previous population estimates for the region, and usable density estimates to extrapolate population sizes are only available for fallow land (Moreira & Leitão 1996b; Moreira 1999), which does not represent the habitat with the highest abundance (see below).

*Figure A.1.11 - Map of Short-toed Lark (*Calandrella brachydactyla*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

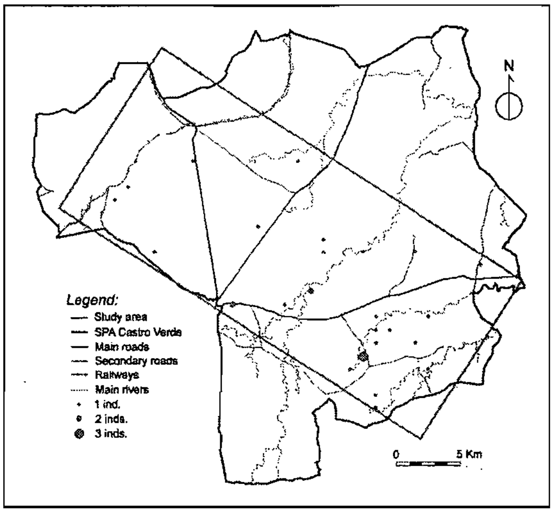


The species was uncommon, although it occurred all over the region (Figure A.1.11). Increasing availability of fallow land and ploughed fields favoured its occurrence (Table A.1.2). Previous studies have shown that this species prefers sparse vegetation including dunes, low density shrublands, fallow fields, and ploughed land (Rufino 1989; Díaz 1994; Suárez *et al.* 2002). In Castro Verde, Delgado & Moreira (Delgado & Moreira 2000) found that Short-toed Larks were most abundant in ploughed land, although they also occurred in fallow fields and pastures. Densities in fallow grasslands increased where fields had a higher proportion of bare ground (Moreira 1999).

- *Montagu's Harrier*

The Montagu's Harrier occurred in 6.6% of the sampling points, with an average abundance of 0.077 birds per point (range = 0-3; SE = 0.016), which yielded an estimate of 872 (95% CI=521-1223) birds for the total pseudo-steppe area of the Castro Verde SPA. Yet, the used methodology is not suitable for accurately censusing raptors, therefore results should be interpreted cautiously. In a previous study, Franco *et al.* (1996) estimated a population density of 45-50 pairs per 10,000 ha in the area of Castro Verde SPA, which would correspond to a population size of ca. 500 individuals. The Portuguese population is estimated as 900-1200 individuals (BirdLife International 2004), which suggests that Castro Verde is one of the strongholds for this species in Portugal. This harrier occurred scattered in the region (Figure A.1.12). It was more likely to be seen in points with higher availability of cereal fields and fallow land with scattered shrub patches (Table A.1.2). These results confirm the preferred habitat types of Montagu's harrier for breeding and feeding (Hagemeijer & Blair 1997; Millon *et al.* 2002) in most of its distribution range. These habitats seem to be the most suitable since they are likely to contain large amounts of arthropods, microtine rodents, and birds, which represent the main prey taken by the species (Hiraldo *et al.* 1975; Arroyo 1998).

*Figure A.1.12 - Map of Montagu's Harrier (*Circus pygargus*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

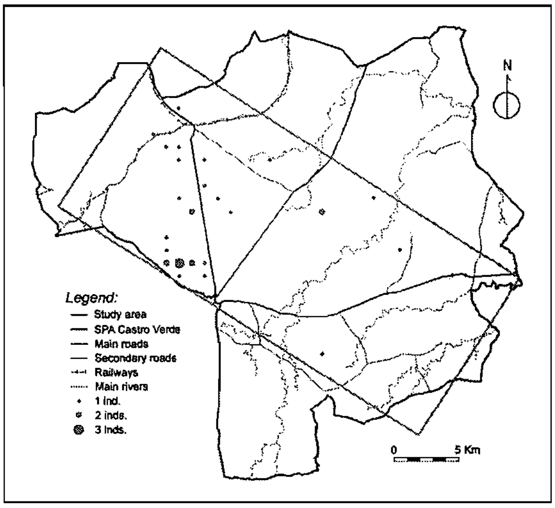


- *Quail*

The Quail was detected in 6.6% of the sampling points, with an average abundance of 0.082 birds per point (range = 0-3; SE = 0.017), which yielded an estimate of 930 (95% CI = 553-1307) individuals for the total pseudo-steppe area of the Castro Verde SPA. BirdLife International (BirdLife International 2004) roughly estimated a national population of 5,000 to 50,000 pairs (based on data from 2002) which suggests that the Castro Verde plains may represent from 1% to 9% of the total Portuguese population.

The species was more common in the western part of the study area (Figure A.1.13), and the likelihood of occurrence increased proportionally with the availability of cereal fields (Table A.1.2). The association of the species with this habitat is consistent with results from previous studies during the breeding season (Carvalho *et al.* 1996; Borralho *et al.* 1998; Delgado & Moreira 2000). In comparison with the Zitting Cisticola, this species responded positively not only to the presence of cereal fields, but to their abundance, suggesting that it may favour larger patches of cereal crops.

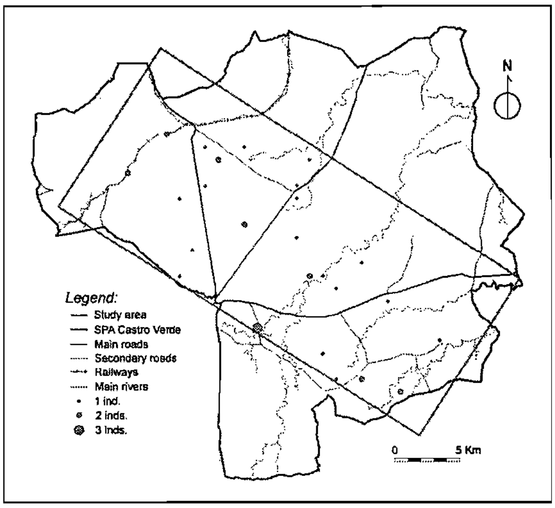
*Figure A.1.13 - Map of Quail (*Coturnix coturnix*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



- *Stone Curlew*

The Stone Curlew was detected in 6.4% of the sampling points, with an average abundance of 0.087 birds per point (range = 0-3; SE = 0.018), which yielded an estimate of 988 (95% CI = 579-1397) individuals for the total pseudo-steppe area of the Castro Verde SPA. The used methodology is not suitable for accurately censusing this species, so care should be taken when interpreting this result. Population estimates for Portugal range from 2500 to 10,000 birds (Cabral *et al.* 2006), thus Castro Verde could hold at least 10%, and up to 40% of the national population. For Castro Verde, the previous estimate of 100-150 pairs (Costa *et al.* 2003) is lower than the current one.

*Figure A.1.14 - Map of Stone Curlew (*Burhinus oedicnemus*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



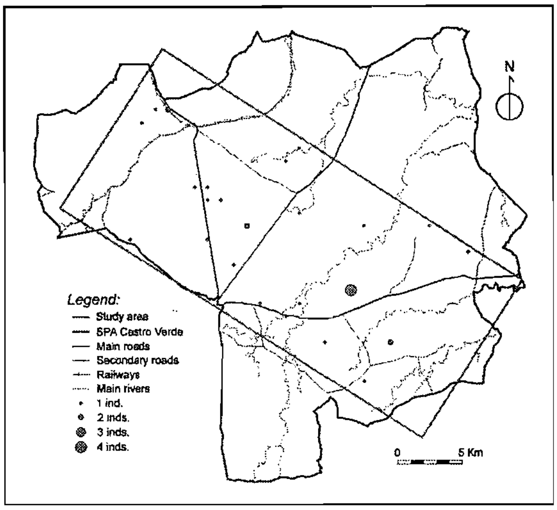
The species occurred scattered over the area (Figure A.1.14), and none of the studied variables influenced its probability of occurrence (Table A.1.2). Moreira (Moreira 1999) found that the presence of the species in fallow land was associated with shrub occurrence, while Delgado & Moreira (Delgado & Moreira 2000) found relatively high densities in ploughed land when compared with other habitats. In the Alto Alentejo, Brito (Brito 1996) also found a significant selection of uncultivated fields with scattered scrubs and a large proportion of bare ground. These results suggest that habitat selection patterns of the Stone Curlew are determined by vegetation structure and soil ground-cover variables that were probably not addressed at the appropriate scale in the present analysis.

- *Hoopoe*

The Hoopoe was present in 6.1% of the sampling points, with an average abundance of 0.077 birds per point (range = 0-4; SE = 0.017), which yielded an estimate of 872 (95% CI=485-1258) individuals for the total pseudo-steppe area of the Castro Verde SPA. The 10,000 to 100,000 pairs estimated by BirdLife International (BirdLife International 2004) suggest that Castro Verde's plains are not very important for this species. This is not surprising if we consider that, in

Iberia, the highest densities occur in open holm oak stands and juniper woodlands (Santos *et al.* 1981{Díaz, 1996 #1266{Muñoz, 2003 #1280).

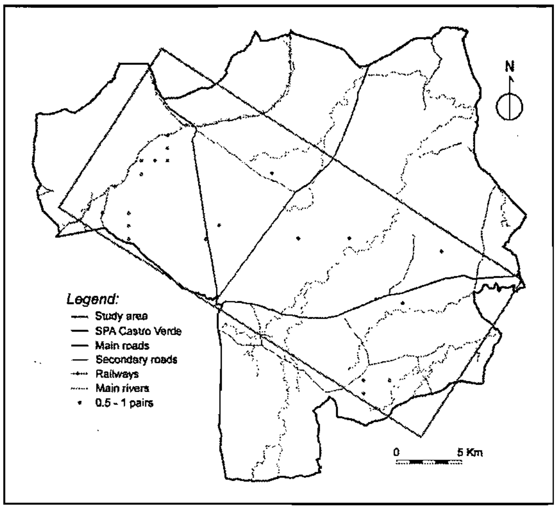
*Figure A.1.15 - Map of Hoopoe (*Upupa epops*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



Hoopoes occurred in a scattered pattern all over the region (Figure A.1.15). The likelihood of finding this species increased proportionally with habitat richness in the points (Table A.1.2). The Hoopoe is basically a bird of warm, dry, level or gently undulating terrain with much exposed bare surface, but numerous upstanding features offering perches, shade and breeding cavities, thus avoiding extensive featureless open tracts, like some large irrigated cultivation and pasture fields (Snow & Perrins 1998). In southern Europe, it is common on farmland with walls and isolated trees, bare or sparsely vegetated soil being in every case essential for ground feeding (Bannerman 1955). Mixed landscapes where woods alternate with cultivation, fallow and pasture fields also appear to be its favourite habitat in Portugal (Rufino 1989), and agrees with our finding of preference for areas with higher habitat diversity. In the most featureless areas of Castro Verde plain pseudo-steppe, the Hoopoe breeds mostly on piles of stones removed from cultivation fields to make ploughs easier, a situation also noted in Spain (Muñoz & Altamirano 2003).

- *Tawny Pipit*

*Figure A.1.16 - Map ofTawny Pipit (*Anthus campestris*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



The Tawny Pipit occurred in 5.4% of the sampling points, with an average abundance of 0.041 birds per point (range = 0-1; SE = 0,009), which yielded an estimate of 465 (95% CI=260-670) pairs for the total pseudo-steppe area of the Castro Verde SPA. The current estimate of the Portuguese population of this species is 1,000-10,000 pairs (BirdLife International 2004), which means that Castro Verde could hold 5 to 45% of the national population.

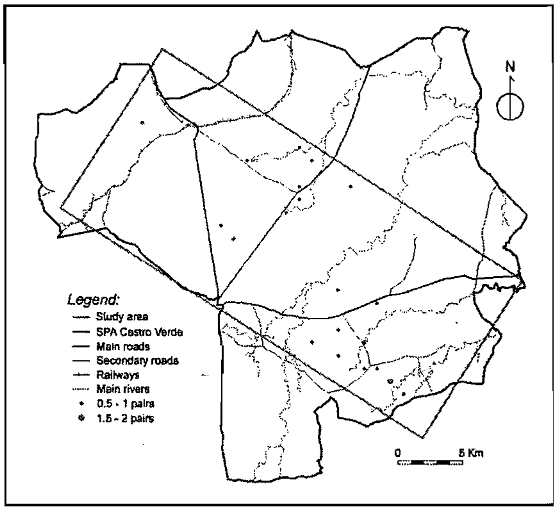
The species was scarce but occurred across the whole region. However, the data suggests that it was more prevalent in the western part (Figure A.1.16). None of the studied variables influenced its probability of occurrence (Table A.1.2). Rufino (Rufino 1989) describes it as a species typical of mountain pastures and also fallow land. Previous studies in Castro Verde found highest densities in ploughed land (Delgado & Moreira 2000).

- *Black-eared Wheatear*

The Black-eared Wheatear occurred in 5.1% of the sampling points, with an average abundance of 0.049 birds per point (range = 0-2; SE = 0.011), which yielded an estimate of 552 (95% CI=302-802) wheatear pairs for the total pseudo-steppe area of the Castro Verde SPA. The national breeding population has been estimated as 2,000-20,000 pairs (BirdLife International 2004), the species being more common and abundant in the south of the country (Rufino 1989). More recently, Almeida *et al.* (2006) reported that the Portuguese breeding population was probably less than 10,000 birds. Considering this estimate, the Castro Verde SPA could support at least 10% of the national breeding population.

The species was very scarce but occurred across the whole region (Figure A.1.17). However, it seemed more prevalent in the eastern part. Its occurrence was positively related to the presence of ploughed fields, and negatively related to the availability of cereal fields (Table A.1.2). In Portugal, this wheatear is strongly associated with very dry land with poor vegetation cover, as fallow land and a variety of habitats with poor or low vegetation cover and height (Rufino 1989). Delgado & Moreira (2000) reported higher densities on ploughed land for Castro Verde. The present distribution in Castro Verde SPA may partially be explained by both habitat and climatic features, as the eastern part of the study area is drier, probably with poor soils and less vegetation cover.

*Figure A.1.17 - Map of Black-eared Wheatear (*Oenanthe hispanica*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

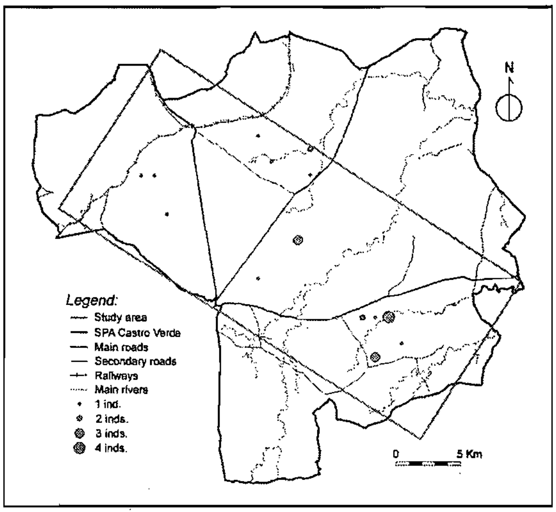


- *Great Bustard*

The Great Bustard was present in 4.1% of the sampling points, with an average abundance of 0.064 birds per point (range = 0-4; SE = 0.018), which yielded an estimate of 726 (95% CI=316-1137) individuals for the total pseudo-steppe area of the Castro Verde SPA. Pinto *et al.* (2005) estimated ca. 900 birds in this SPA, corresponding to 80% of the national population. Thus, although the methodology used was not suitable for accurately censusing this species (Alonso & Alonso 1996; Pinto *et al.* 2005), the population estimate reflects the real population size in the area.

The spatial distribution pattern across the region (Figure A.1.18) partially reflects what is known on the main areas of occurrence during the breeding season (Rocha 1999; Morgado & Moreira 2000), with more observations occurring close to the main lekking grounds. Nevertheless, other known lekking areas were not detected by our sampling scheme.

*Figure A.1.18 - Map of Great Bustard (*Otis tarda*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*

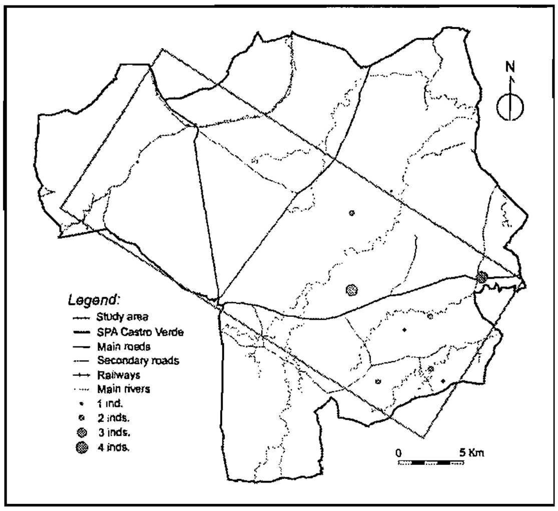


None of the studied variables influenced the probability of occurrence of this species in the sampled points (Table A.1.2), probably because the spatial scale and methods were not suitable for describing habitat selection patterns in the Great Bustard. Previous studies in Castro Verde showed gender differences in habitat selection during the breeding season, with males showing stronger selection for fallows and females preferring cereal fields (Morgado & Moreira 2000; Moreira *et al.* 2004).

- *Black-bellied Sandgrouse*

The Black-bellied Sandgrouse was the scarcest of the studied species, being detected in just 2.0% of the sampling points, and with an average abundance of 0.046 birds per point (range = 0-4; SE = 0.018). This yielded an estimate of 523 (95% CI = 122-924) individuals for the total pseudo-steppe area of the Castro Verde SPA, but the methodology used is not suitable for accurately censusing this species. This probably explains why the obtained estimate was higher than that given by Costa *et al.* (2003) for Castro Verde (20-120 birds). More recently, a Black-bellied Sandgrouse national census estimated that the Portuguese population is not larger than 300 individuals, and counted just 50 individuals in the Castro Verde region, during spring (Cardoso 2005).

*Figure A.1.19 - Map of Blackbellied Sandgrouse (*Pterocles orientalis*) occurrence. Sample points where the species did not occur are not shown, and dot size is proportional to abundance*



The species occurred mainly in the eastern part of the region (Figure A.1.19), a pattern also observed during the national census (Cardoso 2005). Although the area has suitable habitat further west, where observations had occurred in previous years, the present data suggest that the distribution area of Black-bellied Sandgrouse is contracting towards the east (Cardoso 2005).

The probability of finding Black-bellied Sandgrouse was higher in points with a large (over 50%) availability of ploughed land (Table A.1.2). This finding is consistent with a previous study on habitat selection of this species in a nearby area, which found that it prefers areas with scarce vegetation cover (Poeiras 2003). The same study also reports an association with pastures and a high preference for dry leguminous crops.

A.1.5. Discussion

- *Spatial distribution patterns*

The Corn Bunting was substantially more prevalent (almost in 80% of the points) than the other studied species across the region, probably because of a combination of large abundance and high detectability. Little Bustard, Calandra Lark and *Galerida* larks occurred in ca. 30% of the points whereas all the other species had a prevalence value lower than 15%. Most species occurred all over the region, with no obvious spatial pattern of large areas of absence or spatially concentrated occurrence. This could be expected as we focused our effort in the central area of Castro Verde, composed mostly of steppe habitat suitable for the target species. However, areas of higher frequency (and abundance, for Corn Bunting) could be identified, and future monitoring should clarify whether these are determined mostly by annual variations in habitat availability or are consistent across time.

- *Species habitat associations*

The species studied could be grouped into four categories, in terms of relationship with the measured habitat variables: a) species clearly favouring fallow fields and permanent pastures include Calandra Lark and Little Bustard; b) a second category included species associated with cereal fields, namely Quail, Corn Bunting, Zitting Cisticola and Montagu's Harrier; c) a third group was composed of species associated with ploughed fields: Short-toed Lark, Black-eared Wheatear, Black-bellied Sandgrouse; d) a last group included species probably associated to more diverse habitat mosaics or to landscape variables not assessed at the scale used in the present study: *Galerida* larks, Red-legged Partridge, Hoopoe, Stone Curlew, Stonechat and Great Bustard. Some of these species did not show any association with the measured variables.

- *Population estimates*

The population estimates obtained were not corrected for detectability, so they cannot be compared among species and should be considered as an index for future within-species comparisons, assuming detectability remains constant over time (Martin *et al.* 2007). As they stand, they mostly consist of underestimates and, thus, minimum population sizes. Additionally, the estimates available for Portugal are often very crude, hindering the assessment of the national relevance of Castro Verde populations. Even with these constraints, when compared to previous population estimates (Costa *et al.* 2003), the present data show that the importance of Castro Verde for steppe birds is even higher than supposed. For Little Bustard, previous estimates of 360-3340 males (Costa *et al.* 2003; Silva *et al.* 2006) are increased to 3400-5000 males. For Stone Curlew, previous estimates of 100-150 pairs now reached 580-1400 individuals. For Calandra Lark, estimates become more precise, with the minimum rising from 400 pairs (Costa *et al.* 2003) to 4900 pairs and the maximum decreasing from 8500 to 7400 pairs.

This work provides the first population estimates for Corn Bunting, Short-toed Lark, Tawny Pipit and Black-eared Wheatear. For other species, such as Great Bustard, Montagu's Harrier or Black-bellied Sandgrouse, the methodology used cannot be considered appropriate, and other census methods should be used to assess population status.

In terms of national importance, Castro Verde is extremely important for Great Bustard (80% of the Portuguese population), Calandra Lark (over 60%), Little Bustard (20%) and, probably, Montagu's Harrier. Additionally, the region probably holds relevant percentages (10% or more) of the national population of Short-toed Lark, Stone Curlew, Black-eared Wheatear and Black-bellied Sandgrouse. The area is probably also of relevance for the Corn Bunting, although there are no precise estimates of the national population for this species.

Again, we emphasise that this study was carried out in the spring of 2006, following the worst drought of the last 60 years in Portugal. This drought probably had important negative impacts on bird populations, mainly for resident species, which are likely to be reflected in the current results. Even agricultural management practices were changed, for example livestock grazing was introduced in failed cereal crops. Bird censuses at the national level have suggested a strong impact of the drought of 2004/2005 on bird populations (Hilton 2006).

- *Expected trends in habitats and populations – what will happen in the future?*

As a final exercise, based on the obtained species-habitat associations, we hypothesize expected trends in species populations in relation to potential scenarios of land management changes in the region.

With decoupling, dry cereal cultivation will no longer be a profitable option for local farmers. Thus, one likely scenario will be for dry cereal abandonment and its replacement by pastures. This will be detrimental for species associated with cereal fields, such as Corn Bunting, Zitting Cisticola, Quail and Montagu's Harrier. But other species are expected to decline as the end of crop cultivation will probably also mean the end of field ploughing. Thus, species associated with ploughed land, such as the Black-eared Wheatear, Short-toed Lark and Black-bellied Sandgrouse, are also expected to decline.

As another management alternative, in the context of the end of cereal cultivation, afforestations of former agricultural land are increasing in the region. This is also a threat to most steppe birds, mainly species requiring large areas of fallow and pastures such as Calandra Lark and Little Bustard. On the other hand, afforestations could be beneficial for *Galerida* larks, at least in the short to medium term, and for Red-legged Partridges and Corn Buntings, if the long-term consequence is the increase of 'montados' in the area.

On the other hand, the increase of permanent pastures could a priori be considered beneficial for species such as Little Bustard and Calandra Lark, but this will depend on the grazing system, livestock densities and resulting vegetation structure.

Finally, agricultural abandonment and subsequent scrub encroachment are expected, at least in the medium term, to improve habitat suitability for a few species such as *Galerida* larks, and Red-legged Partridges, but would be highly detrimental for typical steppe birds such as Great Bustard, Little Bustard and Calandra Lark.

- *Conclusion*

This study provided the first data on detailed spatial distribution patterns and population estimates for several steppe birds in Castro Verde. The results suggest that the method used is a quick and effective one for characterising occurrence patterns and making population estimates across relatively large areas of pseudo-steppe, as well as describing broad scale bird-habitat relationships. The main value of the data obtained in this project is to use them as a baseline situation against which the results of future monitoring can be compared. We propose that changes in habitat and bird populations should be monitored at least every 5 years.

A.1.6. Acknowledgements

This study was carried out within the scope of the EUFAR (European Fleet for Airborne Research) research project STEPPEBIRD – using multi-scale remote sensing to study habitat selection by cereal steppe birds in Portugal. Thanks are due to Liga para a Protecção da Natureza (LPN), Sociedade Portuguesa para o Estudo das Aves (SPEA) and Parque Natural do Vale do Guadiana (PNVG) for their support to this project, and to the students of Universidade Lusófona de Humanidades e Tecnologias (ULHT) that helped organizing the logistics of the expedition.

A.2. The effects of species and habitat positional errors on the performance and interpretation of species distribution models

A.2.1. Abstract

Aim: A key assumption in species distribution modelling is that both species and environmental data layers contain no positional errors yet this will rarely be true. This study assesses the effect of introduced positional errors on the performance and interpretation of species distribution models.

Location: Baixo Alentejo region of Portugal

Methods: Data on steppe bird occurrence were collected using a random stratified sampling design on a 1 km^2 pixel grid. Environmental data were sourced from satellite imagery and digital maps. Error was deliberately introduced into the species data as shifts in a random direction of 0-1 pixels, 2-3 pixels, 4-5 pixels and 0-5 pixels. Whole habitat layers were shifted by one pixel to cause misregistration and the cumulative effect of one to three shifted layers investigated. Distribution models were built for three species using three algorithms with three replicates. Test models were compared with controls without errors.

Results: Positional errors in the species data led to a drop in model performance, although not enough for models to be rejected. Model interpretation was more severely affected with inconsistencies in the contributing variables. Errors in the habitat layers had similar although lesser effects.

Main conclusions: Models with species positional errors are hard to detect, often statistically good, ecologically plausible and useful for prediction, but interpreting them is dangerous. Misregistered habitat layers produce smaller effects probably because shifting entire layers does not break down the correlation structure to the same extent as random shifts in individual species observations. Spatial autocorrelation in the habitat layers protects against positional errors to some extent but they should be minimised through careful field design and processing.

Keywords: Misregistration; location error; spatial autocorrelation; species distribution model; steppe birds

A.2.2. Introduction

Species distribution modelling (SDM) refers to a group of techniques for predicting the full range of a species over a given geographical area from incomplete species location data (sightings, museum records, radio-tracked fixes etc.). The numerous algorithms available for SDM (e.g. Elith *et al.* 2006) work by relating the presence of a species to environmental conditions at geographic locations where it occurs to form some expression of the ecological niche (Soberón 2007). An implicit assumption in this process is that both the species location and the associated environmental conditions are measured without positional errors, yet there is good reason to suppose that this will not often be true. Attempts to relate species and environmental data with unknown location errors are unlikely to model true ecological relationships and may lead to inappropriate conservation actions (Loiselle *et al.* 2003; Visscher 2006). The problems of positional errors and error propagation in spatial modelling are not new (e.g. Heuvelink 1998) and there have been many attempts to understand and limit the impact of image registration errors on change detection in remote sensing (e.g. Townshend *et al.* 1992; Dai & Khorram 1998; Wang & Ellis 2005) and citing papers). Methods also exist for handling georeferencing errors and calculating uncertainty (Wieczorek *et al.* 2004) although these have not been used in SDM to our knowledge. Despite this background, it is only recently that attention has been paid to positional errors in SDM. Van Niel *et al.* (2004) studied the effects of errors in Digital Elevation Models (DEM) on vegetation modelling, noting that similar problems probably exist for other environmental data. Hines *et al.* (2005) noted the impact of errors in mapped forest characteristics in delineating suitable habitat for spotted owls *Strix occidentalis occidentalis*. Graham *et al.* (2008) have evaluated how errors in the species data may affect SDM with a focus on a comparison of modelling algorithms, while Visscher (Visscher 2006) has considered the implications of error in GPS telemetry data. Only Johnson & Gillingham (Johnson & Gillingham 2008) appear to have considered both errors in species data (Global Positioning System (GPS) locations) and environmental data (misclassification of land classes), although Graham *et al.* (2008) also note that this is important.

In this paper, we extend existing studies by exploring the impact of both typical and extreme positional errors in species and environmental data on the performance of SDM for steppe birds in Portugal. We consider not only metrics of the predictive performance of the models (e.g. Graham *et al.* 2008) but also their ecological interpretation and the spatial pattern in their predictions which could influence conservation practitioners. Scale is likely to be an important determinant of the effect of positional errors in SDM. By typical errors, we mean positional shifts that are small relative to the grain (pixel) size at which analysis is undertaken. Co-registration errors in RS, for example, are likely to be smaller than a single pixel, and modern animal survey data recorded by GPS are likely to be placed in the correct pixel when rasterized down to 100 x 100 m or so. Examining more extreme errors is important, however, both to understand the trends in error propagation as error increases, and in anticipation of a growth in studies using finer resolution satellite imagery as a source of predictors. Recent sensors such as Ikonos, QuickBird, OrbView-3 and GeoEye-1 provide data below 4 x 4 m resolution, heralding the possibility of detailed within-territory habitat assessments for birds and other organisms. With these sensors, it is evident that errors in the species data will often be multiples of the grain size at which analysis is possible.

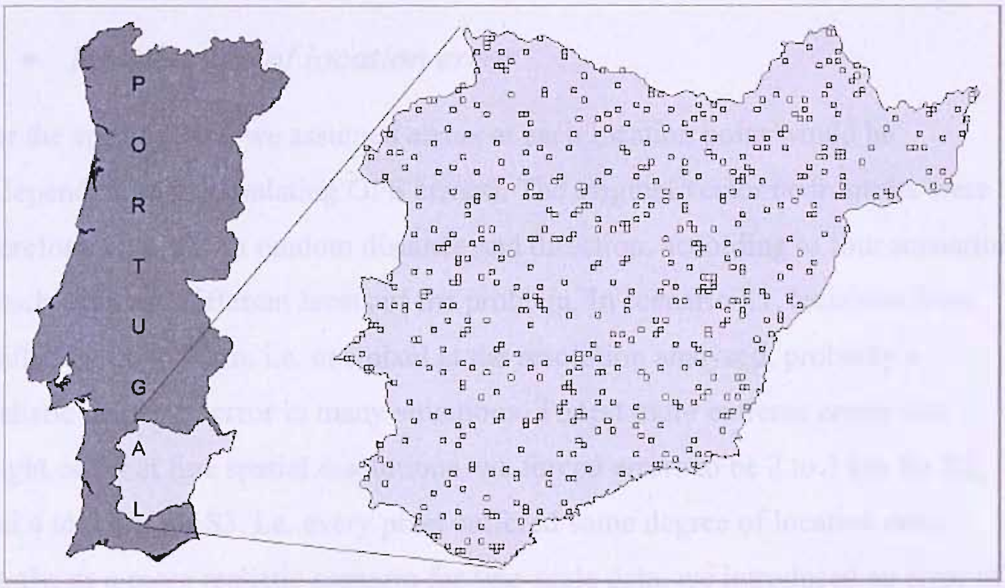
A.2.3. Methods

- *Data sources*

The field data were collected in spring 2004 in the Baixo Alentejo region of Portugal (Figure A.2.1). This region covers ca. 8500 km², comprising a mix of semi-natural Mediterranean habitats including rolling cereal-steppes, fallow fields, open woodlands ("'montados'"), shrublands, olive groves and vineyards. The region is important for steppe birds and includes three Special Protection Areas (SPAs) e.g. Castro Verde SPA, the main steppe area in Portugal and of international importance for several steppe bird species (Moreira *et al.* 2007) (see Appendix A.1). Bird data were gathered systematically in 560 1 km² grid squares according to a stratified random sampling design, close to the ideal for SDM

(Araújo & Guisan 2006). Each square was surveyed once during the early morning or evening for 30 minutes and the GPS locations of breeding species recorded. By using a GPS, we estimate the maximum location error in the species data to be < 100 m, well within the grain size used for analysis (1 km²). For this analysis we chose three species with contrasting habitat preferences: *Melanocorypha calandra* (calandra lark) which favours fallow land; *Circus pygargus* (Montagu's harrier) which prefers cereals; and *Elanus caeruleus* (Black-winged kite) which uses the 'montados'.

Figure A.2.1 - The study area comprised the Baixo Alentejo region of Portugal covering ca. 8,500 km². Bird surveys (white dots) were conducted at 560 locations on a 1 km² grid



The environmental data layers were derived from RS and map data at a spatial resolution of 1 km^2, following the approach of Osborne *et al.* (2001). Vegetation was described by using a 12-month time series of Normalized Difference Vegetation Index (NDVI) images from the SPOT VEGETATION sensor (www.spotvegetation.com/vegetationprogramme/). We calculated monthly images as maximum value composites of three successive 10-day images for each month to minimise the effects of cloud cover and reduce sun-angle, shadow, aerosol and water-vapour effects, all of which can reduce data reliability (Holben 1986). For modelling purposes, we reduced the time-series into seven variables which described vegetation characteristics according to season. Terrain variables were derived from a DEM acquired from the Instituto Geográfico Português (IGP), originally at a spatial resolution of 250 x 250 m. From this we calculated average altitude (ALT) and topographic variability with a 10 m vertical resolution (TOPOV10) within each 1 km^2 grid square (Suárez-Seoane *et al.* 2002a). Proxies for disturbance were calculated as

distance to urbanisation (URBANDIST) derived from the Corine Land Cover 2000 raster data provided by the EEA, and distance to roads (ROADDIST) derived from a vector-based road map provided by the Instituto de Estradas de Portugal (IEP).

- *Introduction of location error*

For the species data, we assumed errors at each location point would be independent (e.g. simulating GPS errors). The original vector coordinates were therefore shifted by a random distance and direction, according to four scenarios which explored different facets of the problem. In scenario S1, locations were shifted by up to 1 km, i.e. one pixel at the resolution analysed, probably a realistic degree of error in many situations. To test more extreme errors that might occur at fine spatial resolutions, we forced errors to be 2 to 3 km for S2, and 4 to 5 km for S3, i.e. every pixel suffered some degree of location error. Lastly, as a more realistic scenario for fine-scale data, we introduced an error of 0 to 5 km for S4. In all cases, values were drawn from a uniform random distribution (see Discussion). After shifting, the data were rasterised to a 1 km^2 grid for analysis. Scenario S1 therefore corresponded to a positional error in the species data of zero or one pixel, S2 two or three pixels, S3 four or five pixels, and S4 zero to five pixels. Three stochastic realisations of each scenario were run to indicate variability.

For the habitat data we assumed that the commonest positional errors would arise from misregistration of GIS or imagery data layers. Errors for each pixel in a single layer would therefore have uniform distance and direction from the true position. We also reasoned that misregistration would arise in the parent data layers such that all derived layers would be similarly affected (van Niel *et al.* 2004). For example, if a DEM and the species data were misregistered, a consistent error would apply to derived variables such as altitude, slope and aspect. In habitat scenario H1, we shifted one parent layer by one pixel in a random direction, varying which layer was chosen on each of three runs. Under scenario H2, a further (different) parent layer was shifted by one pixel and added to H1, such that the three runs in H2 each had two different shifted parent layers. Similarly, under H3 we combined three shifted parent layers. Table A.2.1 shows which parent layers were shifted in each of the three habitat scenarios. As shifting the habitat layers meant they no longer overlapped perfectly, we reduced the size of the study area to the common area across all layers to prevent computational difficulties.

Table A.2.1 - The three scenarios, H1 to H3, with introduced error in the habitat layers. For example, run 2 of scenario H2 had both DEM and NDVI layers shifted while run 1 of H3 had NDVI, the river map and DEM shifted

		Model run	
Scenario code	**1**	**2**	**3**
H1	NDVI imagery	DEM	Road map
H2	+ River map	+ NDVI imagery	+ DEM
H3	+ DEM	+ Land cover map	+ NDVI imagery

- *Modelling*

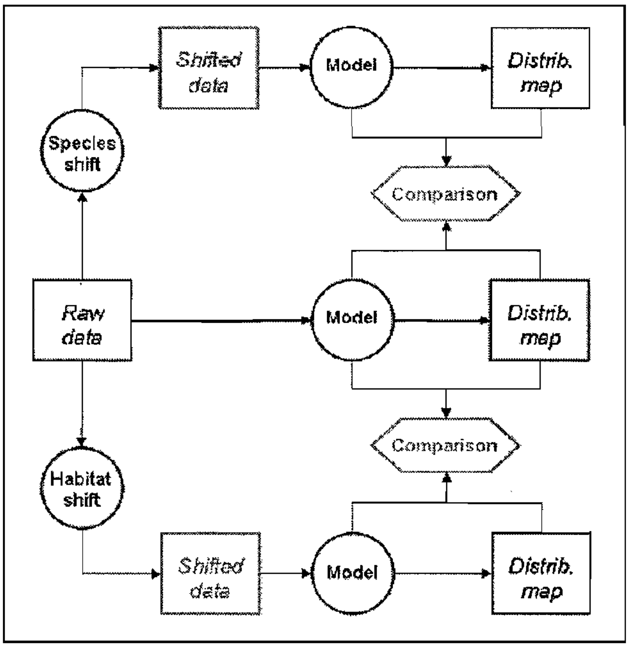
Models were run using two popular and reliable algorithms (Elith *et al.* 2006) with one variant in the model selection routine. Our aim was not to compare the algorithms as such but to assess the way positional errors were propagated through the modelling process to the final output. As an example of a good general technique for presence and verified absence data, we used generalised additive modelling (GAM) (Hastie & Tibshirani 1990) with model selection based on multiple competing hypotheses assessed using Akaike's Information Criterion (AIC) (Akaike 1974; Burnham & Anderson 2002, 2004). As an alternative empirical approach more suited to prediction than interpretation, we used a GAM with an automated backward stepwise selection routine based on AIC (stepGAM) (Lehmann *et al.* 2002). To simulate situations where verified absence data are lacking, Maximum Entropy Modelling (MaxEnt, v.2.1) was used with the presence data contrasted against a randomly drawn sample of background pixels (see Phillips *et al.* 2006) and (Phillips & Dudík 2008) for computational details).

Model fits were assessed using the explained deviance in the GAMs and change in gain within MaxEnt (an analogue of deviance). Performance was evaluated using the area under the receiver operating characteristics curve (AUC) which has comparable although slightly different meaning when applied to GAMs and MaxEnt (see Beck & Shultz 1986; Phillips *et al.* 2006). Comparison of model interpretation within algorithms was made by calculating the drop contribution of each variable to the model on each run of each scenario, drop contributions being assessed by the change in deviance or gain when a variable was removed from the overall model. To quantify differences in mapped outputs from control and scenario models, we first standardised the probability scores to favourability to remove the effects of prevalence (Real *et al.* 2006) and then calculated the mean absolute difference in the predictions across all pixels as:

$$1 - [\Sigma|(Ci - Si)\,|\,/\,p]$$

where Ci is the value of the ith pixel in the control model, Si the corresponding pixel in the scenario model, and p the total number of pixels in the model, using the Map Comparison Kit (Visser & De Nijs 2006). Identical models have a score of 1 using this metric and zero would indicate totally dissimilar predictive maps. Control models without introduced errors were calculated for each algorithm and all results expressed relative to these for ease of comparison (Figure A.2.2).

Figure A.2.2 - Schematic representation of the modelling process



A.2.4. Results

- *Control models*

All three modelling approaches produced "good" models (judged by the AUC statistic: Table A.2.2.) that were ecologically plausible for each species. Models for *Elanus* were weakest whereas those for *Melanocorphya* were strongest according to stepGAM and GAM while MaxEnt produced the best model for *Circus*.

Table A.2.2 - Mean and SE of the change in AUC between the control model and models with introduced error in the species (S1 to S4) or habitat (H1 to H3) data. Values are expressed as a proportion of the AUC achieved for the corresponding control model (third column from left)

Algorithm	Species	Control AUC	AUC for each scenario as proportion of control AUC						
			S1	S2	S3	S4	H1	H2	H3
StepGAM	*Circus*	0.84	0.98±0.011	0.94±0.018	0.91±0.022	0.94±0.012	0.97±0.005	0.97±0.009	0.97±0.008
	Elanus	0.77	0.99±0.039	0.85±0.048	1.01±0.050	0.97±0.028	0.96±0.013	0.96±0.025	0.97±0.024
	Melanocorypha	0.88	0.97±0.002	0.94±0.020	0.85±0.023	0.93±0.016	0.99±0.013	0.97±0.010	0.98±0.013
GAM	*Circus*	0.80	0.97±0.005	0.92±0.004	0.88±0.014	0.93±0.012	0.97±0.009	0.96±0.007	0.96±0.010
	Elanus	0.74	0.99±0.022	0.86±0.023	0.90±0.028	0.91±0.013	0.98±0.013	0.96±0.017	0.94±0.004
	Melanocorypha	0.85	0.99±0.009	0.96±0.011	0.88±0.008	0.93±0.008	0.98±0.003	0.98±0.001	0.98±0.004
MaxEnt	*Circus*	0.88	1.00±0.006	0.96±0.011	0.95±0.008	0.97±0.012	0.99±0.004	0.99±0.005	0.99±0.001
	Elanus	0.76	1.01±0.018	0.95±0.019	1.00±0.029	0.99±0.005	0.98±0.021	0.98±0.024	0.98±0.024
	Melanocorypha	0.85	0.96±0.006	0.95±0.011	0.88±0.006	0.93±0.009	0.99±0.008	0.98±0.010	0.98±0.014

- *Models with error introduced into the species location data*

In the majority of cases (species, algorithms and scenarios), positional error in the species data led to a drop in AUC although in S3 with stepGAM and S1 with MaxEnt, *Elanus* showed a marginally "improved" model (Table A.2.2). The drop in AUC for the 0 to 1 pixel shift (S1) was, however, no more than 4% on average, suggesting that the models would still be regarded as "good" (*sensu* (Hosmer & Lemeshow 2000) when judged by AUC alone. Even shifts of 4 to 5 pixels (S3) failed to reduce AUC in all cases although typically around a 10% average reduction was observed. The standard errors in the drop between runs were highest for *Elanus* which also had the weakest control models, perhaps indicating that overall model quality has an impact on the way error affects performance.

The changes in deviance or gain showed a broadly similar pattern to AUC (Table A.2.3) although differences were more pronounced. In three cases for S1 and one for S3, the models with introduced error were "better" than the control models but in all other cases were weaker. The decline in fit was correlated with the degree of error in Circus and *Melanocorypha* but erratic for *Elanus*. The latter species also showed high standard errors for stepGAM, indicative of different predictor variables being selected on different runs within a scenario.

Table A.2.3 - Mean and SE of the change in deviance (StepGAM and GAM) or gain (MaxEnt) between the control model and models with introduced error in the species (S1 to S4) or habitat (H1 to H3) data. Values are expressed as a proportion of the deviance/gain achieved for the corresponding control model

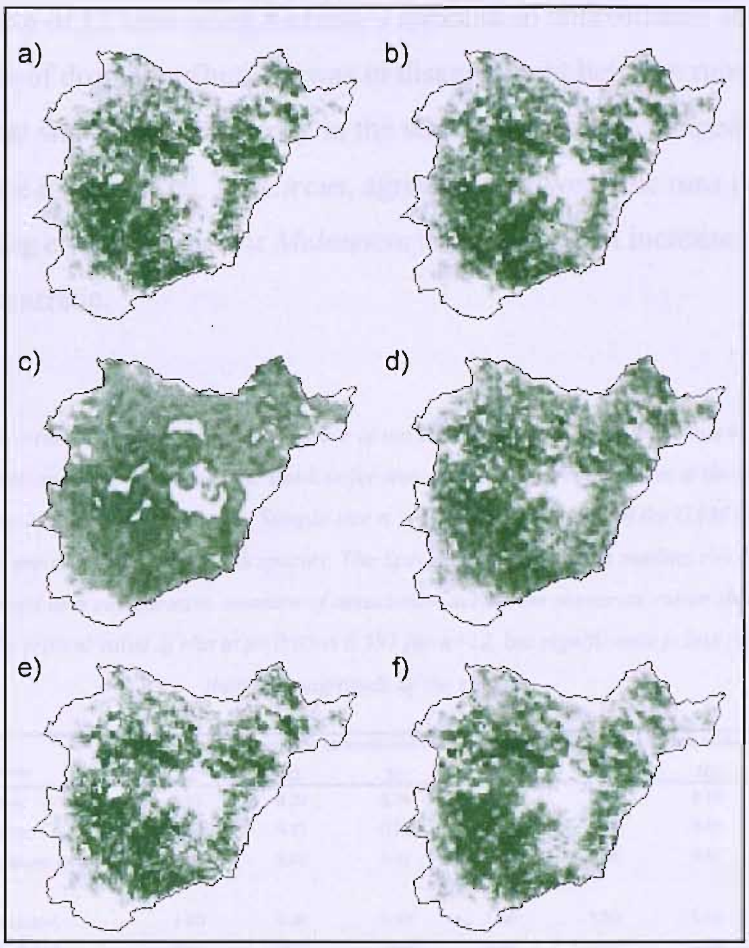
		Deviance/gain as proportion of control for scenario						
Algorithm	Species	S1	S2	S3	S4	H1	H2	H3
StepGAM	*Circus*	0.91±0.049	0.75±0.080	0.59±0.080	0.75±0.063	0.88±0.019	0.87±0.043	0.85±0.038
	Elanus	1.01±0.230	0.39±0.117	1.09±0.306	0.79±0.103	0.75±0.070	0.80±0.102	0.80±0.130
	Melanocorypha	0.82±0.008	0.71±0.067	0.40±0.067	0.64±0.047	0.95±0.081	0.86±0.056	0.92±0.057
GAM	*Circus*	0.84±0.024	0.59±0.023	0.46±0.046	0.66±0.062	0.84±0.043	0.82±0.032	0.78±0.037
	Elanus	0.97±0.123	0.40±0.057	0.47±0.064	0.54±0.066	0.88±0.075	0.76±0.057	0.68±0.008
	Melanocorypha	0.92±0.066	0.78±0.054	0.47±0.036	0.67±0.031	0.91±0.020	0.88±0.010	0.90±0.004
MaxEnt	*Circus*	1.04±0.023	0.80±0.018	0.73±0.056	0.89±0.054	1.04±0.021	1.04±0.025	0.97±0.036
	Elanus	1.23±0.109	0.71±0.037	0.85±0.066	0.85±0.026	1.17±0.094	1.05±0.103	1.03±0.072
	Melanocorypha	0.89±0.055	0.83±0.044	0.52±0.048	0.72±0.050	1.00±0.046	0.96±0.036	0.95±0.040

When we compared the predictive maps produced by the models, the trend in their similarity with the control models was again predictable across scenarios S1 to S3 (Table A.2.4). The effect of S4 (0-5 pixels shift) was always greater than the effect of S1 (0-1 pixel shift). Despite this, the visual change in the models as error was increased was not always apparent (e.g. Figure A.2.3 which shows the stepGAM results for *Melanocorypha*). In general, the core areas were recognised by the error models while peripheral detail varied between scenarios.

Table A.2.4 - Comparison of predictive map outputs from the control models and models with introduced error in the species (S1 to S4) or habitat (H1 to H3) data. The metric shown is the mean and SE of 1 - mean absolute difference in predictions across all map pixels

		1 – mean absolute difference for scenario:						
Algorithm	Species	S1	S2	S3	S4	H1	H2	H3
StepGAM	*Circus*	0.93±0.021	0.89±0.001	0.85±0.020	0.85±0.013	0.94±0.022	0.90±0.013	0.93±0.028
	Elanus	0.91±0.016	0.84±0.010	0.82±0.036	0.85±0.013	0.91±0.010	0.87±0.021	0.87±0.010
	Melanocerypha	0.92±0.004	0.86±0.002	0.85±0.036	0.82±0.004	0.93±0.027	0.88±0.003	0.90±0.034
GAM	*Circus*	0.98±0.005	0.93±0.002	0.91±0.006	0.91±0.015	0.96±0.014	0.94±0.005	0.95±0.014
	Elanus	0.97±0.002	0.88±0.004	0.87±0.022	0.86±0.008	0.95±0.014	0.93±0.006	0.91±0.007
	Melanocorypha	0.97±0.005	0.93±0.002	0.92±0.018	0.89±0.008	0.96±0.017	0.93±0.001	0.94±0.016
MaxEnt	*Circus*	0.88±0.004	0.86±0.009	0.84±0.011	0.81±0.004	0.87±0.006	0.86±0.003	0.86±0.009
	Elanus	0.89±0.004	0.80±0.001	0.80±0.022	0.83±0.018	0.87±0.016	0.84±0.008	0.85±0.022
	Melanocorypha	0.94±0.005	0.92±0.001	0.88±0.022	0.87±0.009	0.93±0.009	0.91±0.004	0.92±0.012

Figure A.2.3 - Examples of output maps from stepGAM models on Melanocorypha (run with median AUC chosen to illustrate each scenario). (a) – control model with no error; (b) S1 (run 2) – 0 to 1 pixel shift; (c) S3 (run 1) – 4 to 5 pixel shift; (d) S4 (run 1) – 0 to 5 pixel shift; (e) H1 (run 2) – 1 pixel shift in one layer; (f) H3 (run 1) – 1 pixel shift in three layers



While the absolute contributions of variables to models with error might not reasonably be expected to be the same as in the control models, the rank order of variables would need to be preserved if ecological interpretations are to be consistent. To test this, we calculated the rank correlation between the drop contributions of variables to the control model with the mean of drop contributions across the three runs for each scenario (Table A.2.5). In general, the small amount of error introduced in S1 had a modest effect on the variable contributions, with MaxEnt suffering worst. The extreme error in S3 often led to major changes in which variables appeared in the models, sometimes with almost no correlation between the drop contributions (e.g. *Melanocorypha* with MaxEnt). This instability in the important predictors probably explains the changing appearance of the output map (Figure A.2.3 – see especially S3). We also noted that the drop contributions were not consistent between the different runs for each scenario. For example, for the three species and four scenarios with MaxEnt, 11 out of 12 tests using Kendall's measure of concordance showed that the rank order of drop contributions was in disagreement between runs (at p<0.05). There was little consistency in the way concordance changed with the degree of error in the model: for *Circus*, agreement between the runs declined with increasing error whereas for *Melanocorypha* agreement increased and for *Elanus* it was erratic.

Table A.2.5 - Correlations between the rank order of variable contributions to the scenario models with introduced error and the control models. Rank order was calculated from the mean of the contributions across the three runs for each scenario. Sample size n = 12 for all cases except the GAM models where samples sizes are given above for each species. The Spearman's correlation statistic rho as used here should be viewed as a comparative measure of association across the scenarios rather than a test of a hypothesis. The critical value of rho at p<0.05 is 0.591 for n=12, but significance is less important here than the magnitude of the statistic

		Correlation between rank order of variables in control model and scenario						
Algorithm	Species	S1	S2	S3	S4	H1	H2	H3
StepGAM	*Circus*	0.83	0.39	0.70	0.71	0.94	0.76	0.69
	Elanus	0.86	0.27	0.32	0.19	0.95	0.83	0.11
	Melanocorypha	0.90	0.66	0.31	0.89	0.90	0.57	0.14
GAM	*Circus n=4*	1.00	0.80	0.80	1.00	1.00	1.00	1.00
	Elanus n=5	0.90	0.60	-0.30	0.60	0.90	0.90	0.70
	Melanocorypha n=5	1.00	0.90	0.90	0.90	1.00	0.90	1.00
MaxEnt	*Circus*	0.60	0.78	0.45	0.53	0.36	0.59	0.71
	Elanus	0.50	-0.05	0.38	0.30	0.53	0.48	0.37
	Melanocorypha	0.73	0.41	-0.06	0.73	0.81	0.74	0.48

- *Models with error introduced into the habitat data*

The misregistration of habitat data layers caused a decline in AUC in all cases but the largest drop across scenarios was only 6% and more usually no more than 3% (Table A.2.2). Again the standard errors for the species with the weakest model (*Elanus*) were greater than for other species and stepGAM under the H3 scenario produced a "better" model than under the less severe H1 and H2 scenarios. Deviance and gain behaved in a broader similar fashion to AUC but all the MaxEnt models for *Elanus* produced a higher gain with the shifted habitat layers than without (Table A.2.3). Differences between scenarios did not follow the expected pattern of larger drops in explained deviance/gain the more error was added, the effects being rather unpredictable. In comparing the output maps, scenario H2 always produced a greater difference from the control than H1, but H3 often led to smaller differences than H2 despite more layers being shifted (Table A.2.4), suggesting that the number of layers misregistered may be less important than their identity. The rank order of variable contributions to the habitat error models often differed widely from the control and generally increased in severity the more layers were shifted (Table A.2.5). For example, the H3 scenario using stepGAM for *Melanocorypha* had only three variables in common with the control out of a total of nine used, although the visual effect on the probability map was not striking (Figure A.2.3). Overall across all comparison statistics used, the impact of misregistered predictor layers was less than an equivalent pixel shift in the species layer.

A.2.5. Discussion

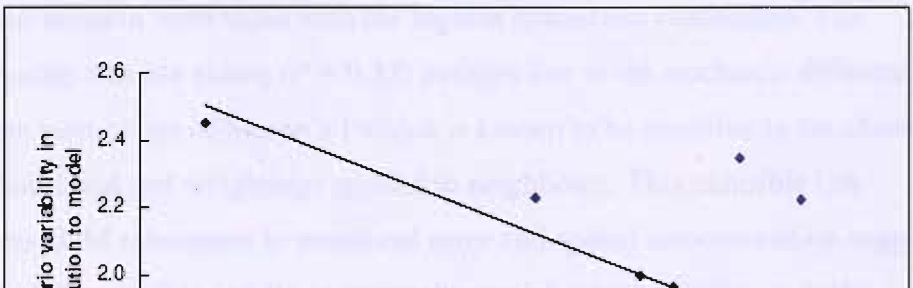
In this study we introduced positional errors into our species data ranging from mild to extreme distortion in order to understand how errors are propagated through the modelling process. Our mildest distortion of up to one pixel (S1) may be the maximum experienced in many studies where the pixel size used for analysis is more than double the likely positional error. It may even be too severe as we drew the random shifts from a uniform distribution (e.g. rather than a normal curve cf. (Johnson & Gillingham 2008) in order to induce more effect (Visscher 2006). It is therefore likely that species positional errors have a small effect on the fit and predictive performance of many models as judged by AUC (concurring with (Graham *et al.* 2008). The corollary of this is that AUC is not a helpful statistic for distinguishing models with and without positional errors and the thresholds for accepting models as "good" based on AUC may be too low (see also (Lobo *et al.* 2008). Changes in deviance or gain for the same models were far more pronounced and in our view provide a better indicator of errors than the small changes in AUC. In the majority of cases with mild errors, however, neither the changes in AUC nor deviance explained were sufficient to cause the researcher to reject the model (according to the thresholds for acceptable models based on AUC – see Swets 1988; Hosmer & Lemeshow 2000; Pearce & Ferrier 2000). In consequence, models containing positional errors have probably been interpreted ecologically and used in applied contexts. Analysis of the predictive maps produced, however, showed strong similarities especially in the core areas of a species range. If our results are typical of other situations, it is unlikely that interpretation of such maps by conservation managers would have led to grave errors in identifying key areas for protection or the boundaries to the main areas used by a species, although they could certainly mislead in more marginal areas. Our findings therefore agree with those of Graham *et al.* (2008) that useful predictions can be made even when species data contain some positional error. In addition, however, our results show that the predictors driving such models were markedly affected by errors (see also (Johnson & Gillingham 2008) whereas Graham *et al.* (2008) did not study model inference. When we examined the variable contributions, we found large inconsistencies between models with introduced error and those without, even between runs of a single scenario. Yet in many cases, it would be possible for a competent ecologist to "explain" why the selected variables were important to the species. Unless positional errors are known to be very small, it is dangerous to infer that the variables selected by the modelling algorithms are those used by the species during habitat selection, or even proxies for them, and their coefficients cannot be taken to indicate importance to the species. Thus although predictive success may be preserved in models built from datasets with positional errors, ecological interpretation is not. Our take-home message is that models with species positional errors are hard to detect, often statistically good, ecologically plausible and useful for prediction, but interpreting them is dangerous. Although it was not our main purpose to compare algorithms, our results are consistent with those of Graham *et al.* (2008) in noting differences between them and that MaxEnt appears robust to moderate levels of error.

The introduction of extreme positional errors (up to 5 pixels) into the species data caused larger changes in fit than the smaller errors: in general, the greater the error, the more AUC and deviance or gain declined. Again, however, AUC was often insufficiently affected for the models to be rejected. We used these extreme errors to indicate what might happen when researchers start to use the increasingly available very high resolution imagery as predictors in their models (i.e. to explore scale effects). Even at a resolution of 10 x 10 m pixels, the errors in species locations are likely to be several pixels unless extreme care is taken (e.g. using differential GPS). Our analysis suggests that models built with large degrees of positional error will be weaker and very often based on different predictors than control models.

Our findings also showed that misregistration of habitat layers caused smaller effects on the models than a shift of the same magnitude in the species data ((Johnson & Gillingham 2008) also found that positional errors in the species data had the largest effect on model outcomes). In fact, it was not always apparent that an increase in the number of layers shifted caused greater impact on the models. Overall the effects were less severe and less predictable than corresponding species positional errors. One reason why this might be true is a difference in the way the errors are propagated. Species errors usually have both random distance and direction: if the predictor data layers have low spatial autocorrelation (see below) the effect of species errors is to break-down the correlation structure with the predictors. In contrast, misregistered habitat layers are often shifted in a single direction by a fixed amount (but see Brown *et al.* 2007) for treatment of spatial variability in misregistration). A constant directional shift weakens but does not break down the correlation structure between the species and predictor layers. The relationships identified may be "wrong" in the sense that the variables involved are not the cues used by the species in habitat selection, but they remain consistent enough to produce predictive models. In fact, we would go further and argue that all the variables commonly used in bird distribution models (and those for many other species) are only proxies for causal variables. Positional errors cause the substitution of another proxy but the predictions remain valid. This is another reason why interpreting 2 ecologically which variables drive the models is inadvisable.

The over-riding recommendation must be that researchers make themselves aware of the many potential sources of positional errors in SDMs and minimise them through careful field design and data processing. As a general rule of thumb, it would be prudent to consider the likely degree of error in each data layer (both predictor and response variables) and, having identified the largest likely error, restrict analysis to grain sizes at least twice this value. For example, in a study using bird data from point counts (sighting error ~ 10 - 30 m), located using a modern GPS unit (inherent error up to 20 m) and modelled against resampled Landsat imagery (error up to 30 m), it would be unwise to base analysis on single Landsat pixels of 30 m. Instead, a more robust analysis would come from using a 3 x 3 matrix of Landsat pixels (= 90 m resolution) at roughly twice the error in the bird data (which equals 20 + 30 m). The data from individual Landsat pixels need not be lost but should be incorporated as a variability measure within the unit of analysis.

Figure A.2.4 - Average within-scenario variability of a predictor's contribution to the model (across all species and scenarios) against the spatial autocorrelation in the predictor (Rook's case Moran's I). Low values on the y-axis indicate greater consistency in the performance of predictor variables when subject to positional errors; such variables tend to have higher spatial autocorrelation. Analysis based on results from MaxEnt



We suspect that a further indication of the potential impact of positional errors on SDMs may be gained by examining spatial autocorrelation, a common feature of predictor variables more often seen as a challenge than an opportunity (Dormann *et al.* 2007). Logically, errors in species coordinates will matter less if the shifted location shares the same environmental features as the true location (i.e. the habitat variables are spatially auto-correlated). In such cases, analysis will still uncover the ecological relationships (niche) which underpin distribution models. Similarly, when habitat layers are misregistered, the correlation structure between the habitat variables will be preserved if there is strong similarity in the values of adjacent pixels. Although theoretically true, we know of no demonstration of this on real ecological data and therefore performed an exploratory test for our study site. If spatial autocorrelation serves to "rescue" models with location errors, we would expect to see a link between this autocorrelation and consistency in the rank position of variables within models across any situation. We calculated the standard deviation in rank position of each variable across the three runs for each scenario (S1 to S4) for each species and then averaged these standard deviations (n = three species x four scenarios). We then plotted this measure against Moran's I for each variable (Figure A.2.4). As surmised, the variables which appeared most consistently in models despite errors in location were those with the highest spatial autocorrelation. The relationship was not strong ($r^2 = 0.37$) perhaps due to the stochastic differences between runs or use of Moran's I which is known to be sensitive to the chosen neighbourhood and weightings applied to neighbours. This plausible link between SDM robustness to positional error and spatial autocorrelation suggests that modellers could usefully examine the spatial autocorrelation in each predictor variable (after the appropriate spatial resolution for analysis has been chosen) to indicate vulnerability. Ironically, the variables with most resilience to positional errors could also be those with poorest predictive power since they would be spatially invariant. There is clearly a trade-off between local spatial invariance to protect against the effects of positional error and the need for larger scale variability to separate good from poor habitat for the species being studied.

A.2.6. Acknowledgements

This work was funded by FCT grant SFRH/BD/12569/2003 with additional support from Liga para a Protecção da Natureza (LPN), Sociedade Portuguesa para o Estudo das Aves (SPEA), Instituto de Estradas de Portugal (IEP) and Perímetro Florestal da Contenda. Supplementary data was provided by Pedro Rocha, Ana Delgado, Inês Henriques and projects EDIA/PMo5.4, PRAXIS/C/AGR/11062/1998, PRAXIS/C/AGR/11063/1998 and LIFE02/NAT/P/8476.

REFERENCES

Adams J. B., Smith M. O. & Johnson P. E. (1986) Spectral mixture modelling: a new analysis of rock and soil types at the Viking Lander 1 site. *Journal of Geophysical Research* 91: 8098-8112.

Akaike H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.

Almeida J. (1992) Census of Common Crane (*Grus grus*) wintering in Portugal. *Airo* 3: 55-58.

Almeida J., Catry P., Encarnação V., Franco C., Granadeiro J. P., Lopes R., Moreira F., Oliveira P., Onofre N., Pacheco C., Pinto M., Pitta Groz M. J., Ramos J. & Silva L. (2006) *Oenanthe hispanica* chasco-ruivo. In: *Livro Vermelho dos Vertebrados de Portugal* (eds. M. J. Cabral, J. Almeida, P. R. Almeida, T. Dellinger, N. Ferrand Almeida, M. E. Oliveira, J. M. Palmeirim, A. I. Queiroz, L. Rogado & M. Santos-Reis) pp. 373-374. Instituto da Conservação da Natureza, Lisboa, Portugal.

Alonso J. C. & Alonso J. A. (1996) The Great Bustard *Otis tarda* in Spain: present status, recent trends and an evaluation of earlier censuses. *Biological Conservation* 77: 79-86.

Alonso J. C., Morales M. B. & Alonso J. A. (2000) Partial migration, and lek and nesting area fidelity in female Great Bustards. *The Condor* 102: 127-136.

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

Araújo M. B., Thuiller W., Williams P. H. & Reginster I. (2005) Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology & Biogeography* 14: 17-30.

Arroyo B. E. (1998) Effect of diet on the reproductive success of Montagu's Harrier *Circus pygargus*. *Ibis* 140: 690-693.

Atkinson P. M. & Tate N. J. (2000) Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer* 52: 607-623.

Augustin N. H., Mugglestone M. A. & Buckland S. T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 33: 339-347.

Austin G. E., Thomas C. J., Houston D. C. & Thompson D. B. A. (1996) Predicting the spatial distribution of buzzard Buteo buteo nesting areas using a geographical information system and remote sensing. *Journal of Applied Ecology* 33: 1541-1550.

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

Austin M. P. (1980) Searching for a model for use in vegetation analysis. *Vegetatio* 42: 11-21.

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

Baldock D. (1991) Implications of EC farming and countryside policies for conservation of lowland dry grassland. In: *The conservation of lowland dry grassland birds in Europe* (eds. P. D. Goriup, L. A. Batten & J. A. Norton) pp. 111-118. Joint Nature Conservation Committee, Peterborough, UK.

Ball G. H. & Hall D. J. (1965) ISODATA: a novel method of data analysis and pattern classification. Stanford Research Institute, Stanford, CA, U.S.A.

Bannerman D. A. (1955) *The Birds of the British Isles – Vol. 4*. Oliver & Boyd, London, U.K.

Barbaro L., Couzi L., Bretagnolle V., Nezan J. & Vetillard F. (2008) Multi-scale habitat selection and foraging ecology of the eurasian hoopoe (Upupa epops) in pine plantations. *Biodiversity and Conservation* 17: 1073-1087.

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

Barry S. C. & Welsh A. H. (2002) Generalized additive modelling and zero inflated count data. *Ecological Modelling* 157: 179-188.

Bayes T. & Price R. (1763) An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. . *Philosophical Transactions* 53: 370-418.

Beaufoy G., Baldock D. & Clark J. (1994) *The nature of farming: low intensity farming systems in nine European countries*. Institute for European Environmental Policy, London, UK.

Beck J. B. & Shultz E. K. (1986) The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology and Laboratory Medicine* 110: 13-20.

Belousov A. I., Verzakov S. A. & von Frese J. (2002) A flexible classification approach with optimal generalisation performance: support vector machines. *Chemometrics and Intelligent Laboratory Systems* 64: 15-25.

Benediktsson J. A., Swain P. H. & Ersoy O. K. (1990) Neural network approaches versus statistical methods on classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 28: 540-551.

Berendse F., Chamberlain D., Kleijn D. & Schekkerman H. (2004) Declining biodiversity in agricultural landscapes and the effectiveness of agri-environment schemes. *Ambio* 33: 499-502.

Bibby C., Burgess N. D., Hill D. A. & Mustoe S. H. (2000) *Bird census techniques - 2nd edition*. Academic Press, London, UK.

Bignal E. M. & McCracken D. I. (1996) Low-intensity farming systems in the conservation of the countryside. *Journal of Applied Ecology* 33: 413-424.

BirdLife International (2004) *Birds in Europe: population estimates, trends and conservation status*. BirdLife International, Cambridge, UK.

Blackburn T. M. & Gaston K. J. (1996) Spatial patterns in the species richness of birds in the New World. *Ecography* 19: 369-376.

Blackburn T. M. & Gaston K. J. (2002) Scale in macroecology. *Global Ecology & Biogeography* 11: 185-189.

Blanche F. G., Legendre P. & Borcard D. (2008) Forward selection of explanatory variables. *Ecology* 89: 2623-2632.

Borralho R., Carvalho S., Rego F. & Vaz Pinto P. (1999) Habitat correlates of the Red-legged Partridge (*Alectoris rufa*) breeding density on Mediterranean farmland. *Revue de Ecologie (La Terre et la Vie)* 54: 59-69.

Borralho R., Rego F. & Vaz Pinto P. (1997) Demographic trends of Red-legged Partridges (*Alectoris rufa*) in southern Portugal after implementation of management actions. *Game and Wildlife* 14: 585-599.

Borralho R., Stoate C. & Araújo M. (2000) Factors affecting the distribution of Red-legged Partridges Alectoris rufa in an agricultural landscape of southern Portugal. *Bird Study* 47: 304-310.

Borralho R., Stoate C., Araújo M., Rito A. & Carvalho S. (1998) Factores ambientais que afectaram a ocorrência primaveril de Codorniz *Coturnix coturnix* no Baixo Alentejo. In: *Actas do Simpósio Sobre Aves Migradoras na Península Ibérica* (eds. L. T. Costa, H. Costa, M. Araújo & M. A. Silva). Sociedade Portuguesa para o Estudo das Aves, Évora, Portugal.

Boser B. E., Guyon I. M. & Vapnik V. N. (1992) A training algorithm for optimal margin classifiers. In: *Fifth Annual ACM Conference on Computational Learning Theory* pp. 144-152, Pittsburgh, PA, USA.

Boyce M. S. (2006) Scale for resource selection functions. *Diversity and Distributions* 12: 269-276.

Bradbury R. B., Hill R. A., Mason D. C., Hinsley S. A., Wilson J. D., Balzter H., Anderson G. Q. A., Whittingham M. J., Davenport I. J. & Bellamy P. E. (2005) Modelling relationships between birds and vegetation structure using airborne LiDAR data: a review with case studies from agricultural and woodland environments. *Ibis* 147: 443-352.

Breiman L. (1996) Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 24: 2350-2383.

Breiman L. (2001) Random Forests. *Machine Learning* 45: 5-32.

Brito J. C., Godinho R., Luís C., Paulo O. S. & Crespo E. G. (1999) Management strategies for conservation of the lizard *Lacerta schreiberi* in Portugal. *Biological Conservation* 89: 311-319.

Brito P. H. (1996) Nest site selection by the Stone Curlew (*Burhinus oedicnemus*) in Southern Portugal. In: *Conservación de las aves estepárias y su habitat. Actas del Simposium para la Conservación de las Aves Estepárias y su Habitat.* (eds. J. F. Gutiérrez & J. Sanz-Zuasti). Unión de Grupos Naturalistas de Castilla Y León, Valladolid, Spain.

Brotons L., Mañosa S. & Estrada J. (2004a) Modelling the effects of irrigation schemes on the distribution of steppe birds in Mediterranean farmland. *Biodiversity and Conservation* 13: 1039-1058.

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

Brown J. H. (1984) On the relationship between abundance and distribution os species. *The American Naturalist* 124: 255-279.

Brown K. M., Foody G. M. & Atkinson P. M. (2007) Modelling geometric and misregistration error in airborne sensor data to enhance change detection. *International Journal of Remote Sensing* 28: 2857-2879.

Buckland S. T., Burnham K. P. & Augustin N. H. (1997) Model selection: an integral part of inference. *Biometrics* 53: 603-618.

Buckland S. T. & Elston D. A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* 30: 478-495.

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

Burfield I. (2005) The conservation status of steppic birds in Europe. In: *Ecology and conservation of steppe-land birds* (eds. G. Bota, M. B. Morales, S. Mañosa & J. Camprodon) pp. 119-139. Lynx Edicions & Centre Tecnològic Forestal de Catalunya, Barcelona, Spain.

Burnham K. P. & Anderson D. R. (2002) *Model selection and multimodel inferences. A practical information-theoretic approach - 2nd ed.* Springer, New York, USA.

Burnham K. P. & Anderson D. R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33: 261-304.

Burroughs P. (1995) Spatial aspects of ecological data. In: *Data analysis in community and landscape ecology* (eds. R. H. G. Jongman, C. J. F. Ter Braak & O. R. F. van Togeren). Cambridge University Press, Cambridge, UK.

Busby J. R. (1991) BIOCLIM - a bioclimatic analysis and prediction system. *Plant Protection Quarterly* 6: 8-9.

Cabeza M., Araújo 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.

Cabral M. J., Almeida J., Almeida P. R., Dellinger T., Ferrand Almeida N., Oliveira M. E., Palmeirim J. M., Queiroz. A. I., Rogado L. & Santos-Reis M. (2006) *Livro Vermelho dos Vertebrados de Portugal.* Instituto da Conservação da Natureza, Lisboa, Portugal.

Campbell J. B. (1996) *Introduction to Remote Sensing, 2nd ed.* Taylor & Francis, London, UK.

Cardoso A. C. (2005) Censo Nacional de Cortiçol-de-barriga-negra e de Cortiçol-de-barriga-branca. Instituto da Conservação da Natureza.

Carlile D. W., Skalski J. R., Batker J. E., Thomas J. M. & Cullinan V. I. (1989) Determination of ecological scale. *Landscape Ecology* 2: 203-213.

Carol C., Zielinski W. J. & Noss R. F. (1999) Using presence-absence data to build and test spatial habitat models for the Fisher in the Klamath region, U.S.A. *Conservation Biology* 13: 1344-1359.

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

Carroll S. S. & Pearson D. L. (1998) The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. *Ecography* 21: 401-414.

Carvalho S., Araújo M., Borralho R. & Stoate C. (1996) Análise multivariada para identificação de guildas de aves num mosaico agrícola do Baixo Alentejo. In: *Actas do I Congresso de Ornitologia da Sociedade Portuguesa para o Estudo das Aves* (eds. J. C. Farinha, J. Almeida & H. Costa). Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Ceballos J. C. & Bottino M. J. (1997) The discrimination of scenes by principal components analysis of multi-spectral imagery. *International Journal of Remote Sensing* 18: 2347-2449.

Chang C.-C. & Lin C.-J. (2001) LIBSVM: a Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chavez Jr. P. S. (1988) An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment* 24: 459-479.

Cobby D. M., Mason D. C. & Davenport I. J. (2001) Image processing of airborne scanning laser altimetry data for improved river flood modelling. *ISPRS Journal of Photogrammetry & Remote Sensing* 56: 121-138.

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational Pshycology Measurements* 20: 37-46.

Cohen W. B. & Goward S. N. (2004) Landsat's role in ecological applications of remote sensing. *BioScience* 54: 535-545.

Collingham Y. C., Wadsworth R. A., Huntley B. & Hulme P. E. (2000) Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology* 37: 13-27.

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

Coppin P., Jonckheere I., Nackaerts K., Muys B. & Lambin E. (2004) Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing* 25: 1565-1596.

Coreau A. & Martin J.-L. (2007) Multi-scale study of bird species distribution and of their response to vegetation change: a Mediterranean example. *Landscape Ecology* 22: 747-764.

Cortes C. & Vapnik V. (1995) Support-Vector Networks. *Machine Learning* 20: 273-297.

Costa L. T., Nunes M., Geraldes P. & Costa H. (2003) *Zonas importantes para as aves em Portugal.* Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Cracknell A. P. (1998) Synergy in remote sensing - what's in a pixel? *International Journal of Remote Sensing* 19: 2025-2047.

Craven P. & Wahba G. (1979) Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31: 377-403.

Crist E. P. & Cicone R. C. (1984) A physically-based transformation of Thematic Mapper data: the TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing* 22: 256-263.

Cumming G. S. (2000) Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography* 27: 425-440.

Curran P. J., Milton E. J., Atkinson P. M. & Foody G. M. (1998) Remote sensing: from data to understanding. In: *Geocomputation: a primer* (ed. P. A. Longley) pp. 33-59. John Wiley & Sons Ltd, Chichester, UK.

Cushman S. A. & McGarigal K. (2002) Hierarchical, Multi-scale decomposition of species-environment relationships. *Landscape Ecology* 17: 637-646.

Dai X. L. & Khorram S. (1998) The effects of image misregistration on the accuracy of remotely sensed change detection. *IEEE Transactions on Geoscience and Remote Sensing* 36: 1566-1577.

Davenport I. J., Bradbury R. B., Anderson G. Q. A., Hayman G. R. F., Krebs J. R., Mason D. C., Wilson J. D. & Veck N. J. (2000) Improving bird population models using airborne remote sensing. *International Journal of Remote Sensing* 21: 2705-2717.

De'ath G. (2002) Multivariate regression tree: a new technique for model species-environment realtionships. *Ecology* 83: 1105-1117.

De la Concha I. (2005) The Common Agricultural Policy and the role of rural development programmes in the conservation of steppe birds. In: *Ecology and conservation of steppe-land birds* (eds. G. Bota, M. B. Morales, S. Mañosa & J. Camprodon) pp. 237-252. Lynx Edicions & Centre Tecnològic Forestal de Catalunya, Barcelona, Spain.

De Veaux R. D., Psichogios D. C. & Ungar L. H. (1993) A comparison of two nonparametric estimation schemes: MARS and neural networks. *Computers & Chemical Engineering* 17: 819-837.

De Veaux R. D. & Ungar L. H. (1994) Multicollinearity: a tale of two non-parametric regressions In: *Selecting Models from Data: Artificial Inteligence and Statistics IV* (eds. P. Cheeseman & R. W. Oldford) pp. 293-302. Springer-Verlag, New York, USA.

Delgado A. & Moreira F. (2000) Bird assemblages of an Iberian cereal steppe. *Agriculture Ecosystems & Environment* 78: 65-76.

Delgado A. & Moreira F. (2002) Do wheat, barley and oats provide similar habitat and food resources for birds in cereal steppes? *Agriculture Ecosystems & Environment* 93: 441-446.

DeSante D. F. (1981) A field test of the variable circular-plot censusing technique in a California coastal scrub breedind bird community. *Studies in Avian Biology* 6: 177-185.

Díaz M. (1994) Short-toed lark *Calandrella brachydactyla*. In: *Birds in Europe: their conservation status* (eds. G. M. Tucker & M. F. Heath). Birdlife International, Cambridge, UK.

Diefenbach D. R., Brauning D. W. & Mattice J. A. (2003) Variability in grassland bird counts related to observer differences and species detection rates. *The Auk* 120: 1168-1179.

Dobson M. C., Ulaby F. T. & Pierce L. E. (1995) Land-cover classification and estimation of terrain attributes using Synthetic Aperture Radar. *Remote Sensing of Environment* 51: 199-214.

Donnell D. J., Buja A. & Stuetzle W. (1994) Analysis of additive dependencies and concurvities using smallest additive principal components. *The Annals of Statistics* 22: 1635-1668.

Dormann C. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology & Biogeography* 16: 129-138.

Dormann C. F., McPherson J. M., Araújo M. B., Bivand R., Bolliger J., Carl G., Davies R. G., A. H., Jetz W., Kissling W. D., Kühn I., Ohlemüller R., Peres-Neto P. R., Reineking B., Schröder B., Schurr F. M. & Wilson R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30: 609-628.

Drake J. M., Randin C. & Guisan A. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology* 43: 424-432.

Eastman J. R., Weigen J., Kyem P. A. K. & Toledano J. (1995) Raster procedures for multi-criteria/multi-objective decisions. *Photogrammetric Engineering & Remote Sensing* 61: 539-547.

Edwards Jr. T. C., Cutlerb D. R., Zimmermann N. E., Geiserd L. & 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.

Eklundh L. & Singh A. (1993) A comparative-analysis of standardized and unstandardized principal component-analysis in remote sensing. *International Journal of Remote Sensing* 14: 1359-1370.

Elith J., Burgman M. & Regan H. M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling* 157: 3131-3329.

Elith J., Ferrier S., Huettmann F. & Leathwick J. (2005) The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling* 186: 280-289.

Elith J., Graham C. H., Anderson R. P., Dudík 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. E., 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.

Elith J. & Leathwick J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* 13: 265-275.

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

ESRI (2004) Arc GIS 9. Environmental Systems Research Institute, Redlands, USA.

Estes L. D., Okin G. S., Mwangi A. G. & Shugart H. H. (2008) Habitat selection by a rare forest antelope: A multi-scale approach combining field data and imagery from three sensors. *Remote Sensing of Environment* 112: 2033-2050.

Faria N. & Rabaça J. E. (2004) Breeding habitat modelling of the Little Bustard *Tetrax tetrax* in the site of community importance of Cabrela (Portugal). *Ardeola* 51: 331-343.

Farina A. (1998) *Principles and methods in landscape ecology*. Chapman & Hall, London, UK.

Ferrier S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* 51: 331-363.

Ferrier S. & Guisan A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43: 393-404.

Figueiras A., Roca-Pardiñas J. & Cadarso-Suárez C. (2005) A bootstrap method to avoid the effect of concurvity in generalised additive models in time series studies of air pollution. *Journal of Epidemiology and Community Health* 59: 881-884.

Foody G. M. (1992) A fuzzy sets approach to the representation of vegetation continua from remotely sensed data: an example from lowland heath. *Photogrammetric Engineering & Remote Sensing* 58: 221-225.

Foody G. M. (1996) Fuzzy modelling of vegetation from remotely sensed imagery. *Ecological Modelling* 85: 3-12.

Foody G. M. (1999) The continuum of classification fuzziness in thematic mapping. *Photogrammetric Engineering & Remote Sensing* 65: 443-451.

Foody G. M. (2002) Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80: 185-201.

Foody G. M. (2004) Spatial nonstationary and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology & Biogeography* 13: 315-320.

Foody G. M. & Curran P. J. (1994) Scale and environmental remote sensing. In: *Environmental Remote Sensing from Regional to Global Scales* (eds. G. M. Foody & P. J. Curran) pp. 223-232. John Wiley & Sons, Chichester, UK.

Foody G. M., Lucas R. M., Curran P. J. & Honzak M. (1997) Non-linear mixture modelling without end-members using an artificial neural network. *International Journal of Remote Sensing* 18: 937-953.

Foody G. M. & Mathur A. (2004) A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing* 42.

Foody G. M. & Mathur A. (2006) The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment* 103: 179-189.

Forman R. T. T. (1995) *Land mosaics: the ecology of landscapes and regions.* Cambridge University Press, Cambridge, UK.

Forman R. T. T. & Godron M. (1986) *Landscape ecology.* John Wiley & Sons, New York, USA.

Forster B. C. & Best P. (1994) Estimation of SPOT P-mode point spread function and derivation of a deconvolution filter. *ISPRS Journal of Photogrammetry & Remote Sensing* 49: 32-42.

Fortuna M. A. (2002) Selección de hábitat de la perdiz roja *Alectoris rufa* en período reproductor en relación con las características del paisaje de un agrosistema de la Mancha (España). *Ardeola* 49: 59-66.

Franco A., Malico I., Martins H. & Sarmento N. (1996) Abundância e reprodução do tartaranhãocaçador (*Circus pygargus* L.) na região de Castro Verde. *Ciência e Natureza* 2: 21-28.

Franco A. M. A., Marques J. T. & Sutherland W. J. (2005) Is nest-site availability limiting Lesser Kestrel populations? A multiple scale approach. *Ibis* 147: 657-666.

Franco A. M. A. & Sutherland W. J. (2004) Modelling the foraging habitat selection of lesser kestrels: conservation implications of European Agricultural Policies. *Biological Conservation* 120: 63-74.

Freedman L. S., Pee D. & Midthune D. N. (1992) The problem of underestimating the residual error variance in forward stepwise regression. *The Statistician* 41: 405-412.

Friedman J. H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics* 19: 1-67.

Friedman J. H. & Silverman B. W. (1989) Flexible parsimonious smoothing and additive modeling. *Technometrics* 31: 3-39.

Fuller R. J. & Langslow D. R. (1984) Estimating numbers of birds by point counts: how long should counts last? *Bird Study* 31: 195-202.

Fuller R. M., Devereux B. J., Gillings S., Amable G. S. & Hill R. A. (2005) Indices of bird-habitat preference from field surveys of birds and remote sensing of land cover: a study of south-eastern England with wider implications for conservation and biodiversity assessment. *Global Ecology & Biogeography* 14: 223-239.

Fuller R. M., Groom G. B., Mugisha S., Ipulet P., Pomeroy D., Katende A., Bailey R. & Ogutu-Ohwayo R. (1998) The integration of field survey and remote sensing for biodiversity assessment: a case study in the tropical forests and wetlands of Sango Bay, Uganda. *Biological Conservation* 86: 379-391.

García J., Suárez-Seoane S., Miguélez D., Osborne P. E. & Zumalacárregui C. (2007) Spatial analysis of habitat quality in a fragmented population of little bustard (Tetrax tetrax): Implications for conservation. *Biological Conservation* 137: 45-56.

Gardner R. H. & Turner M. G. (1991) Future directions in quantitative landscape ecology. In: *Quantitative methods in landscape ecology* (eds. M. G. Turner & R. H. Gardner) pp. 519-525. Springer-Verlag, New York, USA.

Gaston K. J., Blackburn T. M., Greenwood J. J. D., Gregory R. D., Quinn R. M. & Lawton J. H. (2000) Abundance-occupancy relationships. *Journal of Applied Ecology* 37: 39-59.

Genç L., Dewitt B. & Smith S. (2004) Determination of wetland vegetation height with LIDAR. *Turkish Journal of Agriculture and Forestry* 28: 63-71.

Gering J. C., Crist T. O. & Veech J. A. (2003) Additive partitioning of species diversity across multiple spatial scales: implications for regional conservation of biodiversity. *Conservation Biology* 17: 488-499.

Gillespie T. W., Foody G. M., Rocchini D., Giorgi A. P. & Saatchi S. (2008) Measuring and modelling biodiversity from space. *Progress in Physical Geography* 32: 203-221.

Gómez D., Biging G. & Montero J. (2008) Accuracy statistics for judging soft classification. *International Journal of Remote Sensing* 29: 693-709.

Gordon I. J. & Dennis P. (1996) Multiple-scale impacts of large herbivore grazing and biodiversity management in the uplands. In: *The spatial dynamics of biodiversity: towards an understanding of spatial patterns & processes in the landscape* (eds. I. A. Simpson & P. Dennis) pp. 25-32. IALE-UK, Stirling, UK.

Gottschalk T. K., Huettmann F. & Ehlers M. (2005) Thirty years of analysing and modelling avian habitat relationships using satellite imagery data: a review. *International Journal of Remote Sensing* 26: 2631-2656.

Graham C. H., Elith J., Hijmans R. J., Guisan A., Peterson A. T., Loiselle B. A. & Group T. N. P. S. D. W. (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology* 45: 239-247.

Graham C. H., Ferrier S., Huettmann F., Moritz C. & Peterson A. T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* 19: 497-503.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters* 10: 364-374.

Griffiths G. H., Smith J. M., Veitch N. & Aspinall R. (1993) The ecological interpretation of satellite imagery with special reference to bird habitats. In:

Landscape ecology and geographic information systems (eds. R. Haines-Young, D. R. Green & S. H. Cousins) pp. 255-272. Taylor & Francis, London, UK.

Grimmet R. F. A. & Jones T. A. (1989) *Important Bird Areas in Europe.* International Council for Bird Preservation (ICBP), Cambridge, UK.

Guisan A., Edwards J., T.C. & Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157: 89-100.

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

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

Guisan A., Weiss S. B. & Weiss A. D. (1999) GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* 143: 107-122.

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

Hagemeijer W. J. M. & Blair J. B. eds. (1997) *The EBCC atlas of European breeding birds. Their distribution and abundance.* T & A D Poyser, London, UK.

Hanley J. A. & McNeil B. J. (1982) The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 143: 29-36.

Hansen M. C., Defries R. S., Townshend J. R. G. & Sohlberg R. (2000) Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing* 21: 1331-1364.

Hansson L. & Angelstam P. (1991) Landscape ecology as a theoretical basis for nature conservation. *Landscape Ecology* 5: 191-201.

Hastie T., Tibshirani R. & Buja A. (1994) Flexible Discriminant Analysis by optimal scoring. *Journal of the American Statistical Association* 89: 1255-1270.

Hastie T. & Tibshirani R. J. (1996) Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society: Series B* 58: 155-176.

Hastie T. J. & Tibshirani R. (1990) *Generalized Additive Models.* Chapman & Hall, London, UK.

Haubrock S.-N., Chabrillat S., Lemmnitz C. & Kaufmann H. (2008) Surface soil moisture quantification models from reflectance data under field conditions. *International Journal of Remote Sensing* 29: 3-29.

Heikkinen R. K., Luoto M., Virkkala R., Pearson R. G. & Körber J.-H. (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology & Biogeography* 16: 754-763.

Hepinstall J. A. & Sader S. A. (1997) Using Bayesian statistics, Thematic Mapper satellite imagery, and breeding bird survey data to model bird species probability of occurrence in Maine. *Photogrammetric Engineering & Remote Sensing* 63: 1231-1237.

Heuvelink G. (1998) *Error propagation in environmental modelling with GIS.* Taylor and Francis, London, U.K.

Hill M. J. & Donald G. E. (2003) Estimating spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series. *Remote Sensing of the Environment* 84: 367-384.

Hilton G. (2006) Censo de aves comuns em Portugal. Dados preliminares de 2004 e 2005. Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Hines E. M., Franklin J. & Stephenson J. R. (2005) Estimating the effects of map error on habitat delineation for the California spotted owl in Southern California. *Transactions in GIS* 9: 541-559.

Hiraldo F., Fernández F. & Amores F. (1975) Diet of Montagu's Harrier (*Circus pygargus*) in southwestern Spain. *Acta Vertebrata* 2: 25-55.

Hirzel A. & Guisan A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* 157: 331-341.

Hirzel A. H., Hausser J., Chessel D. & Perrin N. (2002) Ecological Niche Factor Analysis: how to compute habitat-suitability maps without absence data. *Ecology* 83: 2027-2036.

Holben B. N. (1986) Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing* 7: 1417-1434.

Hosmer D. W. & Lemeshow S. (2000) *Applied logistic regression, 2nd ed.* Wiley, New York, USA.

Hotelling H. (1957) The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology* 10: 69-79.

Hsu C.-W. & Lin C.-J. (2002) A comparison of methods for multi-class Support Vector Machines. *IEEE Transactions on Neural Networks* 13: 415-425.

Huang C., Davis L. S. & Townshend J. R. G. (2002a) An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* 23: 725-749.

Huang C., Townshend J. R. G., Liang S., Kalluri S. N. V. & DeFries R. S. (2002b) Impact of sensor's point spread function on land cover characterization: assessment and deconvolution. *Remote Sensing of Environment* 80: 203-212.

Imhoff M. L., Sisk T. D., Milne A., Morgan G. & Orr T. (1997) Remotely sensed indicators of habitat heterogeneity: use of Synthetic Aperture Radar in mapping vegetation structure and bird habitat. *Remote Sensing of Environment* 60: 217-227.

INAG (2005) Seca 2005. Relatório de balanço. Instituto da Água, Lisboa, Portugal.

Jakubauskas M. E., Legates D. R. & Kastens J. H. (2001) Harmonic analysis of time-series AVHRR NDVI data. *Photogrammetric Engineering & Remote Sensing* 67: 461-470.

Janz A., van der Linden S., Waske B. & Hostert P. (2007) imageSVM - A user-oriented tool for advanced classification of hyperspectral data using Support Vector Machines. In: *5th EARSeL SIG IS workshop: "Image spectroscopy: inovation in environmental research"*. EARSeL, Oud Sint-Jan, Bruges, Belgium.

Jensen J. R. (1996) *Introductory digital image processing: a remote sensing perspective - 2nd ed.* Prentice Hall, Upper Saddle River, USA.

Jenson S. K. & Watz F. A. (1979) Principal component analysis and canonical analysis in remote sensing. *Photogrammetric Engineering & Remote Sensing* 45: 793-784.

Johnson C. J. & Gillingham M. P. (2008) Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. *Ecological Modelling* 213: 143-155.

Johnston C. A. (1989) Quantitative analysis of ecotones using a geographic information system. *Photogrammetric Engineering & Remote Sensing* 55: 1643-1647.

Johnston C. A. (1998) *Geographic information systems in ecology*. Blackwell Science, Oxford, UK.

Jolliffe I. T. (1982) A note on the use of principal components in regression. *Applied Statistics* 31: 300-303.

Kadmon R., Farber O. & Danin A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* 13: 853-867.

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

Kauth R. J. & Thomas G. S. (1976) The tasseled cap - a graphic description of the spectral-temporal development of agricultural crops as seen in Landsat. In: *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data* pp. 41-51, West Lafayette, Indiana, U.S.A.

Kerr J. T. & Ostrovsky M. (2003) From space to species: ecological applications for remote sensing. *Trends in Ecology and Evolution* 18: 299-305.

Kim K. I., Franz M. O. & Schölkopf B. (2005) Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27: 1351-1366.

King A. W. (1991) Translating models across scales in the landscape. In: *Quantitative methods in landscape ecology* (eds. M. G. Turner & R. H. Gardner) pp. 479-517. Springer-Verlag, New York, USA.

Kleijn D., Berendse F., Smit R., Gilissen N., Smit J., Brak B. & Groeneveld R. (2004) Ecological effectiveness of agri-environment schemes in different agricultural landscapes in the Netherlands. *Conservation Biology* 18: 775-786.

Kleijn D. & Sutherland W. J. (2003) How effective are European agri-environment schemes in conserving and promoting biodiversity? *Journal of Applied Ecology* 40: 947-969.

Koenig W. D. (1999) Spatial autocorrelation of ecological phenomena. *Trends in Ecology and Evolution* 14: 22-26.

Kohonen T. (1990) The self-organizing map. *Proceedings of the IEEE* 78: 1464-1480.

La Sorte F. A. & Hawkins B. A. (2007) Range maps and species richness patterns: errors of commission and estimates of uncertainty. *Ecography* 30: 649-662.

Lane S. J., Alonso J. C. & Martín C. A. (2001) Habitat preferences of Great Bustard *Otis tarda* flocks in the arable steppes of central Spain: are potentially suitable areas unoccupied? *Journal of Applied Ecology* 38: 193-203.

Lasaponara R. (2006) On the use of principal component analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series. *Ecological Modelling* 194: 429-434.

Lawler J. J., White D., Neilson R. P. & Blaustein A. R. (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology* 12: 1568-1584.

Leathwick J. R., Elith J. & Hastie T. (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199: 188-196.

Leathwick J. R., Rowe D., Richardson J., Elith J. & Hastie T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology* 50: 2034-2052.

Legendre P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology* 74: 1659-1673.

Lehmann A., Overton J. M. & Leathwick J. R. (2002) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling* 157: 189-207.

Leitão D. & Costa L. T. (2001) First approach to the study of the non-breeding abundance and habitat use by the Little Bustrad Tetrax tetrax in the lower Tejo grasslands (South Portugal). *Airo* 11: 37-43.

Leitão D. & Moreira F. (1996) Estrutura e composição das comunidades de aves nidificantes na região de Castro Verde. *Ciência e Natureza* 2: 103-107.

Leitão P. J., Milton E. J., Mockridge B., Osborne P. E. & Moreira F. (2007) Pre-processing issues affecting the use of CASI data for steppe bird habitat monitoring and management in southern Portugal. In: *NERC ARSF Workshop*, Leicester, U.K.

Leitão P. J., Morgado R., Delgado A. & Moreira F. (2002) Influence of landscape metrics on bird populations of arable farmland in southern Portugal. In: *Avian landscape ecology: pure and applied issues in the large-scale ecology of birds* (eds. D. Chamberlain & A. Wilson) pp. 318-321. IALE-UK, Norwich, UK.

Leitão P. J., Osborne P. E. & Moreira F. (2006) The use of large-scale remote sensing and map data to determine steppe land bird distributions in Baixo Alentejo, Portugal. *Journal of Ornithology* 147 (Suppl.1): 201-202.

Lennington R. K., Sorensen C. T. & Heydorn R. P. (1984) A mixture model approach for estimating crop areas from Landsat data. *Remote Sensing of Environment* 14: 197-206.

Leung Y., Mei C.-L. & Zhang W.-X. (2000) Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning A* 32: 871-890.

Levin S. A. (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73: 1943-1967.

Leyequien E., Verrelst J., Slot M., Schaepman-Strub G., Heitkönig I. M. A. & Skidmore A. (2007) Capturing the fugitive: Applying remote sensing to terrestrial animal distribution and diversity. *International Journal of Applied Earth Observation and Geoinformation* 9: 1-20.

Limiñana R., Soutullo A., Urios V. & Surroca M. (2006) Vegetation height selection in Montagu's Harriers Circus pygargus breeding in a natural habitat. *Ardea* 94: 280-284.

Lin H.-T., Lin C.-J. & Weng R. C. (2007) A note on Platt's probabilistic outputs for support vector machines. *Machine Learning* 68: 267-276.

Liu C., Berry P. M., Dawson T. P. & Pearson R. G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385-393.

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

Loiselle B. A., Howell C. A., Graham C. H., Goerck J. M., Brooks T., Smith K. G. & Williams P. H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* 17: 1591-1600.

Loiselle B. A., Jørgensen P. M., Consiglio T., Jiménez I., Blake J. G., Lohmann L. G. & Montiel O. M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography* 35: 105-116.

Longley P. A. (1998) Foundations. In: *Geocomputation: a primer* (eds. P. A. Longley, S. M. Brooks, R. McDonnell & B. MacMillan) pp. 3-15. John Wiley & Sons Ltd, Chichester, UK.

Lu D. & Weng Q. (2007) A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28: 823-870.

Luoto M., Kuussaari M. & Toivonen T. (2002) Modelling butterfly distribution based on remote sensing data. *Journal of Biogeography* 29: 1027-1037.

Maas H.-G. & Vosselman G. (1999) Two algorithms for extracting building models from raw laser altimetry data. *ISPRS Journal of Photogrammetry & Remote Sensing* 54: 153-163.

MacNally R. (2000) Regression and model-building in conservation biology, biogeography and ecology: The distinction between - and reconciliation of - 'predictive' and 'explanatory' models. *Biodiversity and Conservation* 9: 655-671.

MacNally R. (2002) Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodiversity and Conservation* 11: 1397-1401.

MacNally R. & Walsh C. J. (2004) Hierarchical partitioning public-domain software. *Biodiversity and Conservation* 13: 659-660.

Maisongrande P., Duchemin B. & Dedieu G. (2004) VEGETATION/SPOT: an operational mission for the Earth monitoring; presentation of new standard products. *International Journal of Remote Sensing* 25: 9-14.

Manrique J. & Yanes M. (1994) Thekla Lark *Galerida theklae*. In: *Birds in Europe: their conservation status* (eds. G. M. Tucker & M. F. Heath). BirdLife International, Cambridge, U.K.

Martín C. A., Alonso J. C., Alonso J., Pitra C. & Lieckfeldt D. (2002) Great bustard population structure in central Spain: concordant results from genetic analysis and dispersal study *Proceedings of the Royal Society of London B* 269: 119-125.

Martin J., Kitchens W. M. & Hines J. E. (2007) Importance of well-designed monitoring programs for the conservation of endangered species: case study of the Snail Kite. *Conservation Biology* 21: 472-481.

Martínez C. (1994) Habitat selection by the little bustard Tetrax tetrax in cultivated areas of Central Spain. *Biological Conservation* 67: 125-128.

Martinez F. J. & Purroy F. J. (1993) Avifauna reproductora en los sistemas esteparizados ibericos. *Ecologia* 7: 391-401.

Martinez J.-J. & Wool D. (2006) Sampling bias in roadsides: the case of galling aphids on Pistacia trees. *Biodiversity and Conservation* 15: 2109-2121.

Mason D. C., Anderson G. Q. A., Bradbury R. B., Cobby D. M., Davenport I. J., Vandepoll M. & Wilson J. D. (2003) Measurement of habitat predictor variables for organism-habitat models using remote sensing and image segmentation. *International Journal of Remote Sensing* 24: 2515-2532.

Massy W. F. (1965) Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* 60: 234-256.

McCullagh P. & Nelder J. A. (1989) *Generalized Linear Models, 2nd ed.* Chapman & Hall, London, UK.

McCulloch W. S. & Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5: 115-133.

McDermid G. J., Franklin S. E. & LeDrew E. F. (2005) Remote sensing for large-area habitat mapping. *Progress in Physical Geography* 29: 449-474.

McGarigal K. & McComb W. C. (1995) Relationships between landscape structure and breeding birds in the Oregon Coast Range. *Ecological Monographs* 65: 235-260.

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

McPherson J. M., Jetz W. & Rogers D. J. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions-possibilities and limitations. *Ecological Modelling* 192: 499-522.

Meyer C. B. & Thuiller W. (2006) Accuracy of resource selection functions across spatial scales. *Diversity and Distributions* 12: 288-297.

Millon A., Bourrioux J.-L., Riols C. & Bretagnolle V. (2002) Comparative breeding biology of Hen Harrier and Montagu's Harrier: an 8-year study in north-eastern France. *Ibis* 144: 94-105.

Mimmack G. M., Mason S. J. & Galpin J. S. (2001) Choice of distance matrices in cluster analysis: defining regions. *Journal of Climate* 14: 2790–2797.

Moisen G. G. & Frescino T. S. (2002) Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling* 157: 209-225.

Morales M. B., Suárez F. & García de La Morena E. (2006) Response of steppe birds to various levels of farming intensity and of modification of the agricultural landscape: a comparative analysis of their effects on population density and habitat selection in the Little and Great Bustards (Tetrax tetrax and Otis tarda). *Revue d'Ecologie: La Terre et la Vie* 61: 261-270.

Moreira F. (1999) Relationships between vegetation structure and breeding bird densities in fallow cereal steppes in Castro Verde, Portugal. *Bird Study* 46: 309-318.

Moreira F., Beja P., Morgado R., Reino L., Gordinho L., Delgado A. & Borralho R. (2005) Effects of field management and landscape context on grassland wintering birds in Southern Portugal. *Agriculture Ecosystems & Environment* 109: 59-74.

Moreira F. & Leitão D. (1996a) A comunidade de aves nidificantes nos pousios da região de Castro Verde. *Ciência e Natureza* 2: 109-113.

Moreira F. & Leitão D. (1996b) A preliminary study on the breeding bird community in fallows of cereal steppes in Southern Portugal. *Bird Conservation International* 6: 255-259.

Moreira F., Leitão P. J., Morgado R., Alcazar R., Cardoso A., Carrapato C., Delgado A., Geraldes P., Gordinho L., Henriques I., Lecoq M., Leitão D., Marques A. T., Pedroso R., Prego I., Reino L., Rocha P., Tomé R. & Osborne P. E. (2007) Spatial distribution patterns, habitat correlates and population estimates of steppe birds in Castro Verde. *Airo* 17: 5-30.

Moreira F., Morgado R. & Arthur S. (2004) Great bustard Otis tarda habitat selection in relation to agricultural use in southern Portugal. *Wildlife Biology* 10: 251-260.

Morgado R. & Moreira F. (2000) Seasonal population dynamics, nest site selection, sex-ratio and clutch size of the Great Bustard Otis tarda in two adjacent lekking areas. *Ardeola* 47: 237-246.

Morlini I. (2006) On multicollinearity and concurvity in some nonlinear multivariate models. *Statistical Methods and Applications* 15: 3-26.

Muñoz A. R. & Altamirano M. (2003) Abubilla *Upupa epops* In: *Atlas de las Aves Reproductoras de España* (eds. R. Martí & J. C. del Moral) pp. 348-349. Dirección General de Conservación de la Naturaleza – Sociedad Española de Ornitología, Madrid, Spain.

Nagendra H. (2001) Using remote sensing to assess biodiversity. *International Journal of Remote Sensing* 22: 2377-2400.

Neves R. (1998) Breve análise histórica sobre a evolução da paisagem. In: *Atlas das aves invernantes do Baixo Alentejo* (eds. G. L. Elias, L. M. Reino, J. P. Silva, R. Tomé & P. Geraldes) pp. 28-32. Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Nielsen S. E., Johnson C. J., Heard D. C. & Boyce M. S. (2005) Can models of presence-absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography* 28: 197-208.

O'Neill R. V., Johnson A. R. & King A. W. (1989) A hierarchical framework for the analysis of scale. *Landscape Ecology* 3: 193-205.

Oindo B. O., Skidmore A. K. & De Salvo P. (2003) Mapping habitat and biological diversity in the Maasai Mara ecosystem. *International Journal of Remote Sensing* 24: 1053-1069.

Olden J. D. (2003) A species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology* 17: 854-863.

Oñate J. J. (2005) A Reformed CAP? Opportunities and threats for the conservation of steppe-birds and the agri-environment. In: *Ecology and conservation of steppe-land birds* (eds. G. Bota, M. B. Morales, S. Mañosa & J. Camprodon) pp. 253-281. Lynx Edicions & Centre Tecnològic Forestal de Catalunya, Barcelona, Spain.

Oparin M. L. (2008) Recent fauna of ground-nesting birds in Transvolga steppes and Its dynamics in the 20th century. *Biology Bulletin* 35: 422-427.

Opdam P. (1991) Metapopulation theory and habitat fragmentation: a review of holarctic breeding bird studies. *Landscape Ecology* 5: 93-106.

Osborne P. E. (2005) Using GIS, remote sensing and modern statistics to study steppe birds at large spatial scales: a short review essay. In: *Ecology and Conservation of Steppe-land Birds* (eds. G. Bota, M. B. Morales, S. Mañosa & J. Camprodon) pp. 169-184. Lynx Edicions & Centre Tecnològic Forestal de Catalunya, Barcelona.

Osborne P. E., Alonso J. C. & Bryant R. G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology* 38: 458-471.

Osborne P. E., Foody G. M. & Suárez-Seoane S. (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions* 13: 313-323.

Osborne P. E. & Leitão P. J. (*in press*) Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions.*

Osborne P. E. & Suárez-Seoane S. (2002) Should data be partitoned spatially before building large-scale distribution models? *Ecological Modelling* 157: 249-259.

Osborne P. E. & Suárez-Seoane S. (*2007*) Identifying core areas in a species' range using temporal suitability analysis: an example using little bustards *Tetrax tetrax* L. in Spain. *Biodiversity and Conservation* 16: 3505-3518.

Palmeirim J. M. (1988) Automatic mapping of avian species habitat using satellite imagery. *Oikos* 52: 59-68.

Panigrahy S. & Sharma S. A. (1997) Mapping of crop rotation using multidate Indian remote sensing satellite digital data. *ISPRS Journal of Photogrammetry & Remote Sensing* 52: 85-91.

Paolini L., Grings F., Sobrino J. A., Muñoz J. C. J. & Karszenbaum H. (2006) Radiometric correction effects in Landsat multi-date/multi-sensor change detection studies. *International Journal of Remote Sensing* 27: 685-704.

Patterson M. W. & Yool S. R. (1998) Mapping fire-induced vegetation mortality using Landsat Thematic Mapper data: a comparison of linear transformation techniques. *Remote Sensing of Environment* 65: 132-142.

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

Pearce J., Ferrier S. & Scotts D. (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *Journal of Environmental Management* 62: 171-184.

Pearson K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2: 559-572.

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

Phillips S. J., Anderson R. P. & Schapire R. E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231-259.

Phillips S. J. & Dudík M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31: 161-175.

Pienkowski M. W., Stroud D. A. & Bignal E. M. (1989) Estimating bird numbers and distributions in extensive survey areas using remote sensing. In: *Bird Census and Atlas Studies, Proceedings of the XIth International Conference on Bird Census and Atlas Work* (eds. K. Šťastný & V. Bejček) pp. 5-15, Prague.

Pinto M., Rocha P. & Moreira F. (2005) Long-term trends in great bustard (Otis tarda) populations in Portugal suggest concentration in single high quality area. *Biological Conservation* 124: 415-423.

Platt J. C. (2000) Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. In: *Advances in large margin classifiers* (eds. A. J. Smola, P. Bartlett, B. Schölkopf & D. Schuurmans) pp. 61-74. MIT Press, Cambridge, Massachussets, USA.

Poeiras A. S. (2003) Selecção de habitat do Cortiçol-debarriga-preta (*Pterocles orientalis*) no Parque Natural do Vale do Guadiana. Relatório de Estágio da Licenciatura em Biologia. Universidade de Évora, Évora, Portugal.

Poiani K. A., Richter B. D., Anderson M. G. & Richter H. E. (2000) Biodiversity conservation at multiple scales: functional sites, landscapes, and networks. *BioScience* 50: 133-146.

Polasky S., Camm J. D., Solow A. R., Csuti B., White D. & Ding R. (2000) Choosing reserve networks with incomplete species information. *Biological Conservation* 94: 1-10.

Potts J. M. & Elith J. (2006) Comparing species abundance models. *Ecological Modelling* 199: 153-163.

Pressey R. L., Humphries C. J., Margules C. R., Vane-Wright R. I. & Williams P. H. (1993) Beyond opportunism: key principles for systematic reserve selection. *Trends in Ecology and Evolution* 8: 124-128.

Priestnall G., Jaafar J. & Duncan A. (2000) Extracting urban features from LiDAR digital surface models. *Computers, Environment and Urban Systems* 24: 65-78.

Qi J. & Wallace O. (2002) Biophysical attributes estimation from satellite images in arid regions. In: *IEEE International Geoscience and Remote Sensing Symposium* pp. 2000-2002, Toronto, Canada.

Quattrochi D. A. & Pelletier R. E. (1991) Remote sensing for analysis of landscapes: an introduction. In: *Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity* (eds. M. G. Turner & R. H. Gardner) pp. 51-76. Springer-Verlag, New York, USA.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ralph C. J. (1985) Habitat association patterns of forest and steppe birds of northern Patagonia, Argentina. *The Condor* 87: 471-483.

Real R., Barbosa A. M. & Vargas J. M. (2006) Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics* 13: 237-245.

Reddy S. & Dávalos L. M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* 30: 1719-1727.

Reed B. C., Brown J. F., Vanderzee D., Loveland T. R., Merchant J. W. & Ohlen D. O. (1994) Measuring phenological variability from satellite imagery. *Journal of Vegetation Science* 5: 703-714.

Reino L., Beja P., Osborne P. E., Morgado R., Fabião A. & Rotenberry J. T. (*in press*) Distance to edges, edge contrast and landscape fragmentation: interactions affecting farmland birds around forest plantations. *Biological Conservation.*

Ritchie J. C. (1996) Remote sensing applications to hydrology: airborne laser altimeters. *Hydrological Sciences* 41: 625-636.

Robson N. (1997) The evolution of the Common Agricultural Policy and the incorporation of environmental considerations. In: *Farming and birds in Europe. The Common Agricultural Policy and its implications for bird conservation* (eds. D. J. Pain & M. W. Pienkowski) pp. 43-78. Academic Press, San Diego, USA.

Rocha P. (1999) A interpretação ecológica de imagens de satélite e a utilização de Sistemas de Informação Geográfica aplicados à conservação da abetarda *Otis tarda* no Biótopo Corine de Castro Verde. Tese de Mestrado em Gestão de

Recursos Naturais. In: *Instituto Superior de Agronomia*. Universidade Técnica de Lisboa, Lisboa, Portugal.

Rocha P., Araújo A. & Cruz C. (1996) A evolução das populações portuguesas do francelho-das-torres *Falco naumanni*. In: *Actas do I Congresso de Ornitologia da SPEA* (eds. J. C. Farinha, J. Almeida & H. Costa) pp. 97-98. Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Rogers A. S. & Kearney M. S. (2004) Reducing signature variability in unmixing coastal marsh Thematic Mapper scenes using spectral indices. *International Journal of Remote Sensing* 25: 2317-2335.

Romo H., García-Barros E. & Lobo J. M. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography* 29: 873-885.

Rondeaux G., Steven M. & Baret F. (1996) Optimization of Soil-Adjusted Vegetation Indices. *Remote Sensing of Environment* 55: 95-107.

Rosalino L. M., Santos M. J., Beier P. & Santos-Reis M. (2008) Eurasian badger habitat selection in Mediterranean environments: does scale really matter? *Mammalian Biology* 73: 189-198.

Rossi R. E., Mulla D. J., Journel A. G. & Franz E. H. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs* 62: 277-317.

Rouse J. W., Jr., Hass R. H., Schell J. A. & Deering D. H. (1973) Monitoring vegetation systems in the Great Plains with ERTS. In: *Third ERTS Symposium* pp. 309-317. NASA Goddard Space Flight Center, Greenbelt, MD, U.S.A.

Rufino R. (1989) *Atlas das Aves que nidificam em Portugal Continental*. Centro de Estudos de Migrações e Protecção de Aves, Serviço Nacional de Parques Reservas e Conservação da Natureza, Lisboa, Portugal.

Santisteban A. & Muñoz L. (1978) Principal components of a multispectral image: application to a geological problem. *IBM Journal of Research and Development* 22: 444-454.

Santos C. P. (2000) Succession of breeding bird communities after the abandonment of agricultural fields in south-east Portugal. *Ardeola* 47: 171-181.

Santos T. & Suárez F. (2005) Biogeography and population trends of iberian steppe birds. In: *Ecology and conservation of steppe-land birds* (eds. G. Bota, M. B. Morales, S. Mañosa & J. Camprodon) pp. 69-102. Lynx Edicions & Centre Tecnològic Forestal de Caţalunya, Barcelona, Spain.

Santos T., Suárez F. & Tellería J. L. (1981) The bird communities of Iberian Juniper woodlands (*Juniperus thurifera* L.). In: *Censos de aves en el Mediterrâneo* (ed. F. J. Purroy) pp. 79-88 Universidad de Léon, Léon, Spain.

Schaefer J. A. & Mayor S. J. (2007) Geostatistics reveal the scale of habitat selection. *Ecological Modelling* 209: 401-406.

Schölkopf B., Smola A. & K.-L. M. (1997) Kernel principal component analysis. In: *Artificial Neural Networks - ICANN '97: 7th International Conference, Lausanne, Switzerland, October 1997, Proceedings* (eds. W. Gerstner, A. Germond, M. Hasler & J.-D. Nicoud) pp. 583-588. Springer, Berlin, Germany.

Segurado P. & Araújo M. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography* 31: 1555-1568.

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

Seoane J., Justribó J. H., García F., Retamar J., Radabán C. & Atienza J. C. (2006) Habitat-suitability modelling to assess the effects of land-use changes on Dupont's lark Chersophilus duponti: a case study in the Layna Important Bird Area. *Biological Conservation* 128: 241-252.

Settle J. J. & Drake N. A. (1993) Linear mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing* 14: 1159-1177.

Silva J. P., Faria N. & Catry T. (2007) Summer habitat selection and abundance of the threatened little bustard in Iberian agricultural landscapes. *Biological Conservation* 139: 186-194.

Silva J. P., Leitão D., Santos E., Moreira F., Prego I., Pinto M., Lecoq M., Catry T. & Pedroso R. (2006) Preliminary results of Little Bustard census in Alentejo (Portugal). In: *Bustard Conservation in Europe on the last 15 years* (eds. D. Leitão, C. Jolivet, M. Rodriguez & J. P. Tavares). Royal Society for the Protection of Birds, Sandy, U.K.

Silva J. P., Pinto M. & Palmeirim J. M. (2004) Managing landscapes for the Little Bustard *Tetrax tetrax*: lessons from the study of winter habitat selection. *Biological Conservation* 117: 521-528.

Smith M. O., Ustin S. L., Adams J. B. & Gillespie A. R. (1990) Vegetation in deserts: I. A regional measure of abundance from multispectral images. *Remote Sensing of Environment* 31: 1-26.

Snee R. D. & Marquardt D. W. (1984) Collinearity diagnostics depend on the domain of prediction, the model, and the data. *The American Statistician* 38: 83-87.

Snow D. W. & Perrins C. M. (1998) *The Complete Birds of the Western Palearctic CD-ROM – Version 1.0.* Oxford University Press & Optimedia Software, Oxford, U.K.

Soares P. (1999) Cartaxo-comum *Saxicola torquata.* In: *Atlas das Aves Invernantes do Baixo Alentejo* (eds. G. L. Elias, L. M. Reino, T. Silva, R. Tomé & P. Geraldes) pp. 302-303. Sociedade Portuguesa para o Estudo das Aves, Lisboa, Portugal.

Soberón J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters* 10: 1115-1123.

Sousa J. P., da Gama M. M., Pinto P., Keating A., Calhôa F., Lemos M., Castro C., Luz T., Leitão P. & Dias S. (2004) Effects of land-use on Collembola diversity patterns in a Mediterranean landscape. *Pedobiologia* 48: 609-622.

Steinwart I. (2003) On the optimal parameter choice for nu-Support Vector Machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25: 1274-1284.

Stoate C., Borralho R. J. & Araújo M. (2000) Factors affecting Corn Bunting *Miliaria calandra* abundance in a Portuguese agricultural landscape. *Agriculture Ecosystems & Environment* 77: 219-226.

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

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

Stow D. A. (1993) The role of geographic information systems for landscape ecological studies. In: *Landscape ecology and geographic information systems* (eds. R. Haines-Young, D. R. Green & S. H. Cousins) pp. 11-21. Taylor & Francis, London, UK.

Streutker D. R. & Glenn N. F. (2006) LiDAR measurement of sagebrush steppe vegetation heights. *Remote Sensing of Environment* 102: 135-145.

Suárez-Seoane S., Osborne P. E. & Alonso J. C. (2002a) 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.

Suárez-Seoane S., Osborne P. E. & Baudry J. (2002b) Responses of birds of different biogeographic origins and habitat requirements to agricultural land abandonment in northern Spain. *Biological Conservation* 105: 333-344.

Suárez-Seoane S., Osborne P. E. & Rosema A. (2004) Can climate data from METEOSAT improve wildlife distribution models? *Ecography* 27: 629-636.

Suárez F., Garza V. & Morales M. B. (2002) Habitat use of two sibling species, the Short-toed *Calandrella brachydactyla* and the Lesser Short-toed *C. rufescens* Larks, in mainland Spain. *Ardeola* 49: 259-272.

Suárez F., Naveso M. A. & De Juana E. (1997) Farming in the drylands of Spain: birds of the pseudosteppes. In: *Farming and birds in Europe. The Common Agricultural Policy and its implications for bird conservation* (eds. D. J. Pain & M. W. Pienkowski) pp. 297-330. Academic Press, San Diego, USA.

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

Tan S. S. & Smeins F. E. (1996) Predicting grassland community changes with an artificial neural network model. *Ecological Modelling* 84: 91-97.

Tellería J. L., Santos T., Álvarez G. & Saéz-Royuela C. (1988) Avifauna de los campos de cereales del interior de España. In: *Aves de los medios urbano y agricola en las mesetas españolas* (ed. F. Bernis) pp. 173-317. SEO, Madrid, Spain.

ter Braak C. J. F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167--1179.

Termansen M., McClean C. J. & Preston C. D. (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling* 192: 410-424.

Townshend J. R. G. (1981) The spatial resolving power of earth resources satellites. *Progress in Physical Geography* 5: 32-55.

Townshend J. R. G., Huang C., Kalluri S. N. V., Defries R. S., Liang S. & Yang K. (2000) Beware of per-pixel characterization of land cover. *International Journal of Remote Sensing* 21: 839-843.

Townshend J. R. G., Justice C. O., Gurney C. & McManus J. (1992) The impact of misregistration on change detection. *IEEE Transactions on Geoscience and Remote Sensing* 30: 1054-1060.

Traba J., García de la Morena E. L., Morales M. B. & Suárez F. (2007) Determining high value areas for steppe birds in Spain: hot spots, complementarity and the effciency of protected areas. *Biodiversity and Conservation* 16: 3255-3275.

Traba J., Morales M. B., García de la Morena E., Delgado M.-P. & Krištín A. (2008) Selection of breeding territory by little bustard (Tetrax tetrax) males in Central Spain: the role of arthropod availability. *Ecological Research* 23: 615-622.

Tucker C. J. (1979) Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8: 127-150.

Tucker G. M. (1991) The status of lowland dry grassland birds in Europe. In: *The conservation of lowland dry grassland birds in Europe - Proceedings of an international seminar held at the University of Reading 20-22 March 1991* (eds. P. D. Goriup, L. A. Batten & J. A. Norton) pp. 37-48. Joint Nature Conservation Committee (JNCC), Newbury, UK.

Tucker G. M. (1997) Priorities for bird conservation in Europe: the importance of the farmed landscape. In: *Farming and birds in Europe. The Common Agricultural Policy and its implications for bird conservation* (eds. D. J. Pain & M. W. Pienkowski) pp. 79-116. Academic Press, San Diego, USA.

Tucker G. M. & Heath M. F. (1994) *Birds in Europe: their conservation status.* BirdLife International, Cambridge, UK.

Tucker K., Rushton S. P., Sanderson R. A., Martin E. B. & Blaiklock J. (1997) Modelling bird distributions - a combined GIS and Bayesan rule-based approach. *Landscape Ecology* 12: 77-93.

Turner M. G., Dale V. H. & Gardner R. H. (1989) Predicting across scales: theory development and testing. *Landscape Ecology* 3: 245-252.

van Heezik Y. & Seddon P. (1999) Effects of season and habitat on bird abundance and diversity in a steppe desert, northern Saudi Arabia. *Journal of Arid Environments* 43: 301-317.

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

Vayssières M. P., Plant R. E. & Allen-Diaz B. H. (2000) Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science* 11: 679-694.

Vincent P. J. & Haworth J. M. (1983) Poisson regression models of species abundance. *Journal of Biogeography* 10: 153-160.

Visscher D. R. (2006) GPS measurement error and resource selection functions in a fragmented landscape. *Ecography* 29: 458-464.

Visser H. & De Nijs T. (2006) The Map Comparison Kit. *Environmental Modelling & Software* 21: 346-358.

Wallin D. O., Elliott C. C. H., Shugart H. H., Tucker C. J. & Wilhelmi F. (1992) Satellite remote sensing of breeding habitat for an African weaver-bird. *Landscape Ecology* 7: 87-99.

Wang H. Q. & Ellis E. C. (2005) Image misregistration error in change measurements. *Photogrammetric Engineering & Remote Sensing* 71: 1037-1044.

Whittaker R. J., Willis K. J. & Field R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography* 28: 453-470.

Whittingham M. J. (2007) Will agri-environment schemes deliver substantial biodiversity gain, and if not why not? *Journal of Applied Ecology* 44: 1-5.

Whittingham M. J., Swetnam R. D., Wilson J. D., Chamberlain D. E. & Freckleton R. P. (2005) Habitat selection by yellowhammers Emberiza citrinella on lowland farmland at two spatial scales: implications for conservation management. *Journal of Applied Ecology* 42: 270-280.

Wieczorek J., Guo Q. G. & Hijmans R. J. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18: 145-767.

Wiens J. A. (1989) Spatial scaling in ecology. *Functional Ecology* 3: 385-397.

Wilson E. H. & Sader S. A. (2002) Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment* 80: 386-396.

Wilson K. A., Westphal M. I., Possingham H. P. & Elith J. (2005) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* 122: 99-112.

Wilson R. J., Thomas C. D., Fox R., Roy D. B. & Kunin W. E. (2004) Spatial patterns in species distributions reveal biodiversity change. *Nature* 432: 393-396.

Wintle B. A., Elith J. & Potts J. M. (2005) Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology* 30: 719-738.

Wisz M. S., Hijmans R. J., Peterson A. T., Graham C. H., Guisan A. & NCEAS Predicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14: 763-773.

With K. A. & Crist T. O. (1996) Translating across scales: simulating species distributions as the aggregate response of individuals to heterogeneity. *Ecological Modelling* 93: 125-137.