KU LEUVEN

FACULTY OF SCIENCE

Combining climate and remote sensing data in species distribution models

Jorne Biccler

Supervisor: Prof. B. Sandel
Aarhus University

Supervisor: Prof. T. Verdonck
KU Leuven

Co-supervisor: Prof. J.C. Svenning
Aarhus University

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics

Academic year 2015-2016

© Copyright by KU Leuven

Without written permission of the promotors and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculteit Wetenschappen, Geel Huis, Kasteelpark Arenberg 11 bus 2100, 3001 Leuven (Heverlee), Telephone +32 16 32 14 01.

A written permission of the promotor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Preface

Lastly I would like to thank some people. First of all I am grateful to my supervisor Brody whose very relaxed laissez-faire approach to supervising motivated me to try things that I would probably not have considered otherwise [1]. Furthermore, I would like to thank Tim who made it possible for me to go to Denmark and agreed to act as co-examiner. Finally, I would like to thank my floormates who had to deal with me deciding to make our living room my office every once in a while and my occasional grumpy moods when I found out that my code contained bugs.

[1] Apparently semi-supervised learning sounds fancy but doesn't really work that well for SDM.

Summary

. . . something text
more text

Contents

Todo list

Chapter 1

Introduction

Species distribution modelling (SDM)[1] concerns the practice of modelling the distribution of a species by use of explanatory variables. Applications include predicting the effect of climate change (Pearson and Dawson 2003; Pearson, Dawson, and Liu 2004), predicting the impact of invasive species (Strubbe and Matthysen 2008), predicting the occurrence of wildfires (Parisien and Moritz 2009), ...

Some fundamental ecological concepts are introduced in Chapter 2. Firstly, the concept of an ecological niche is introduced and the connection with SDMs is made. Secondly, to enhance the understanding of the niche concept some of the assumptions underlying the practice of SDM are discussed.

The data-sets and variables that are used in this thesis are described in Chapter 3. These variables describe either the climate, human influence, elevation, ... at a certain location. Climate data is nearly always used to model the occurrence probability of species. When the goal of a study is to obtain coarse grain predictions over a large spatial extent the use of climate data is certainly justified by ecological theory (Pearson and Dawson 2003). However, if there is interest in predictions over a relatively small extent, e.g. when selecting the location of a new national park, fine grain remote sensing data might be useful to distinguish between suitable and unsuitable habitat.

Chapter 4 introduces a number of modelling techniques that are often used to model the distribution of species. In Section 4.2 we focus on classical binary classification methods. These methods are directly applicable if there is access to presence-absence data. However, often the data-sets only includes occurrence locations, for example data-sets from natural history musea or citizen science projects are usually of this type. To use presence-only data the classical classification algorithms from Section 4.2 can be adapted, this is done in Section 4.3.2. Another approach is to use one of the algorithms specifically constructed for presence-only data, one of these is introduced in Section 4.3.3.

In practice a large part of constructing SDMs consists of variable selection. The goal of this thesis is to investigate the performance of multiple model selection methods. An overview of methods used to deal with large magnitudes of predictors is given in Chapter 5. More particularly, we will introduce:

[1] The abbreviation SDM will be used for both the verb, species distribution modelling, and the noun, species distribution model.

- Dimensionality reduction of the explanatory variables.
- Regularization.
- Step-wise selection.

Although we will introduce the most important concepts and some applications of species distribution modelling, it is not the goal of this thesis to describe every aspect in detail. Instead we refer to Franklin and Miller 2009 who gave an overview of the field. Other introductory material includes Elith and Leathwick 2009; Guisan and Thuiller 2005; Guisan and Zimmermann 2000. An introduction to most of the statistical methodology can be found in Hastie, Tibshirani, and Friedman 2009.

Chapter 2

The ecological niche concept

2.1 Introduction

In this chapter the concept of the niche of a species is introduced. The ecological niche of a species can, non-rigorously, be defined as the set of environmental conditions where its reproduction rate is larger than or equal to its mortality rate. Although we will speak of the ecological niche, there are in fact at least three different "definitions" that are often used: the Grinnellian niche, the Eltonian niche, and the Hutchinsonian niche. Only a sketch of the niche concept will be given in this section. For a more rigorous description we refer the interested reader to Soberón 2007; Soberón and Nakamura 2009.

2.2 Ecological versus geographical space

In most databases that contain data about species only the location of a presence or absence record is available. Hence, these databases include information about the occurrences or absences in the so-called geographical space. Usually the range of a species' distribution is determined by environmental conditions. We will say that the corresponding variables span the environmental space. It is clear that for each point in the geographical space there is a point in the environmental space. This relation between environmental and geographical space is often called Hutchinson's duality (Colwell and Rangel 2009). A graphical representation of this relation is given in Figure 2.1. This duality relation is fundamental in SDMs, namely the predictors included in the model are usually assumed to be direct or indirect measures of the variables that span the environmental space. Once a model in the environmental space is constructed the duality relation allows us to make maps of the distribution in the geographical space.

In practice a species will often not occur in certain parts of its niche. This can happen because of limited dispersal capabilities of the species, biotic interactions, etc. Such incomplete occupation of the niche leads to the concepts of a fundamental niche and the realized niche. The fundamental niche does not take into account whether or not the species is present, it only represents the suitable conditions. The realized niche is the subset of the fundamental niche where the species is present. These two concepts are depicted in Figure 2.2.

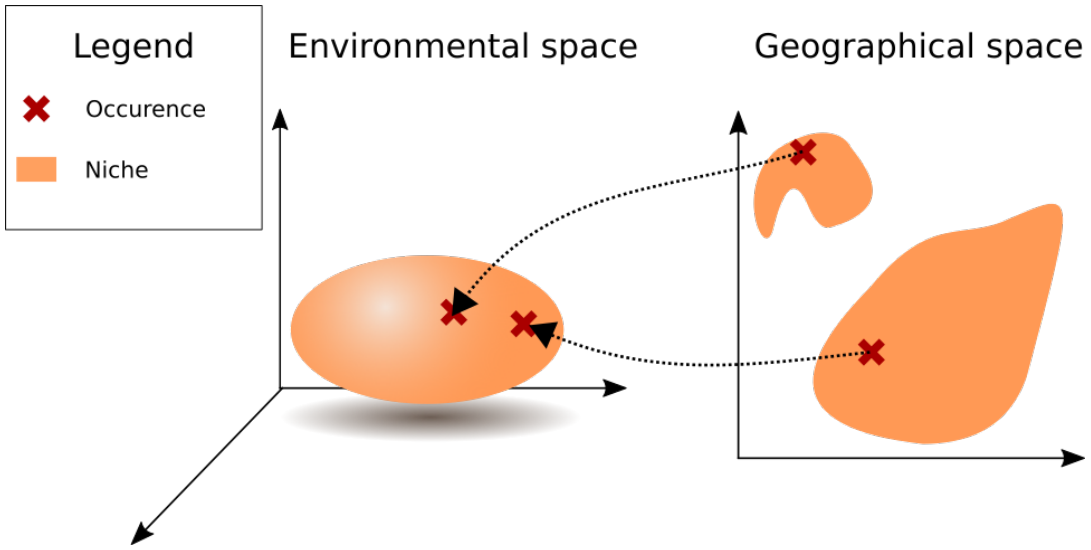

Figure 2.1: Visualization of the duality between environmental and geographical space.

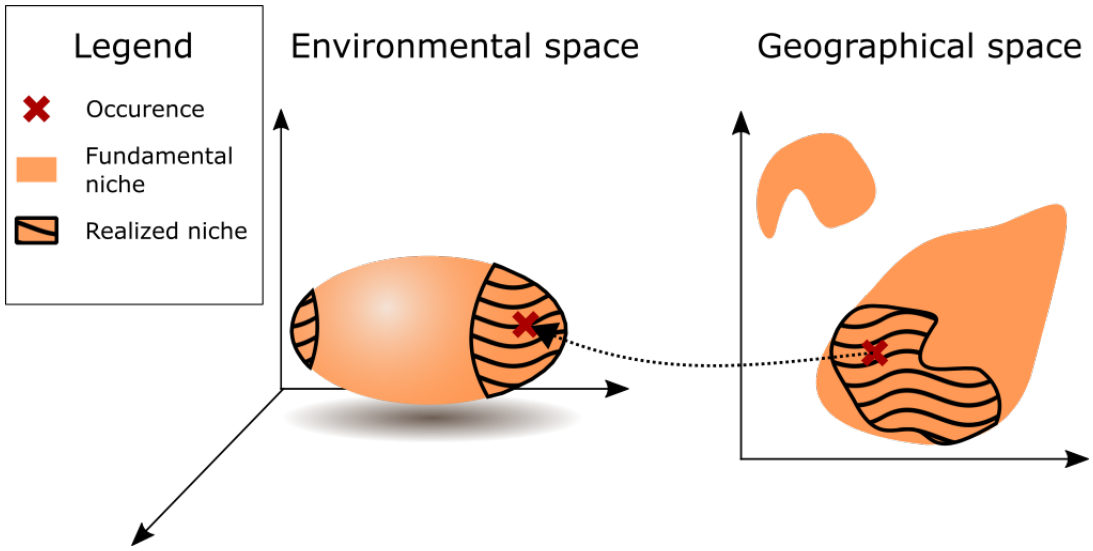

Figure 2.2: Visualization of the difference between the fundamental and realized niche.

2.3 Implicit assumptions when building and using species distribution modles

Before applying SDMs in practice one has to realize that there are some important underlying assumptions. A few of these assumptions are described below. This is done in order to connect the theoretical niche concept to some more practical scenarios and to make the reader aware of the limitations of species distribution modelling. For a more complete overview of the underlying assumptions we refer to Wiens et al. 2009.

Every observed occurrence belongs, by definition, to the realized niche. SDMs are therefore models of the realized niche. If a SDM is used to e.g. predict areas prone to invasive species, it is implicitly assumed that, a part of, the realized niche is a good approximation of, a part of, the fundamental niche. Whether this assumption is realistic or not depends on: the species, whether the whole niche has to be approximated or only a part thereof, etc.

When data is used to build a model of the realized niche it is assumed that the observed data-points are representative of the niche. In practical settings this is often not the case. We give two examples:

1. Due to climate change tree species might be found in regions where the current environmental conditions are not included in its niche (Woodward, Fogg, and Heber 1990).

2. The niche of a species usually evolves over time, therefore older observations might not be representative of the current niche (Pearson and Dawson 2003).

Another assumption that is implicitly made in most SDMs is that the effect of biotic interactions is negligible or indirectly captured by other environmental variables. However, in some applications explicitly including biotic interactions has been shown to improve the predictive capabilities (e.g. Heikkinen et al. 2007).

2.4 Spatial scale

From ecological theory we learn that the spatial scale and the extent of the study area play an important role in species distribution modelling (Pearson and Dawson 2003; Pearson, Dawson, and Liu 2004).

It is often argued that at large scales climate is the main driver of the distribution of species. Land-cover, soil type, and biotic interactions usually only become important when the climate is suitable and hence play a role in fine grain discrimination between suitable and unsuitable conditions. Furthermore, the role of the fine grain drivers might be influenced by the local climatic conditions.

The effect of the extent of the study area and the spatial scale are related. When a SDM is applied to a whole continent climate variables will make the biggest contribution to the model. If on the other hand a SDM is developed for a 10×10 km area in Silkeborg the vegetation type might be more important than minor changes in climatic conditions.

Chapter 3

Data

3.1 Introduction

To study the performance of the different methods we need data-sets containing on explanatory and outcome data. The different explanatory variables that are used are described in Section 3.2. Given the goals of the thesis we opted for using 33 different variables. This is more than used in the average SDM and it should be expected that some of them are quite redundant for the species of interest. The species of interest and data-sets that consist of locations where the species was observed to be present or absent are introduced in Section 3.3.

To process the spatial data R (R Core Team 2013) is used as a geographic information system (GIS). To do this we rely on the RASTER (Hijmans 2015) and SP (Pebesma et al. 2015) packages.

3.2 Predictor data

3.2.1 Vegetation Continuous Fields

The Vegetation Continuous Fields (VCF; DiMiceli et al.) data-set contains values between 0 and 100 which are proportional estimates of the tree cover in the cell. Some cells also contain values larger than 100 which indicate that the cell contains water or that no data is available. Since R has a NA value all the cells with values larger than 100 are set to NA. The raster is provided in the geographical coordinate system combined with the World Geodetic System 1984 datum (GCS_WGS84). The resolution of the VCF raster is 0.00208 decimal degrees.

3.2.2 Bioclimatic variables

The bioclimatic variables are a set of variables that describe ecologically relevant climate patterns.

find a citation

The definition of each of the 19 bioclimatic variables can be found in Table 3.1. The bioclimatic variables can be derived from monthly minimum, maximum, and average temperature and precipitation data. Furthermore, it is clear that the BIO05, BIO6, and BIO7 variables are linearly dependent. This linear dependence can be problematic

when using classification methods. It is interesting that for some of the methods that we introduce this will lead to no problems while it will for others, see Section 5.6.

Variable name	Definition
BIO1	Annual mean temperature
BIO2	Mean diurnal range (mean of monthly $(\text{temp}_{max} - \text{temp}_{min})$)
BIO3	Isothermality $\left(100 \times \frac{\text{BIO2}}{\text{BIO7}}\right)$
BIO4	Temperature seasonality $\left(SD(\text{temp}_{avg}) \times 100\right)$
BIO5	Max temperature of warmest month
BIO6	Min temperature of coldest month
BIO7	Temperature annual range (BIO5 − BIO6)
BIO8	Mean temperature of wettest quarter
BIO9	Mean temperature of driest quarter
BIO10	Mean temperature of warmest quarter
BIO11	Mean temperature of coldest quarter
BIO12	Annual precipitation
BIO13	Precipitation of wettest month
BIO14	Precipitation of driest month
BIO15	Precipitation seasonality $\left(\frac{SD(\text{percipitation})}{\text{mean(percipitaiton)}}\right)$
BIO16	Precipitation of wettest quarter
BIO17	Precipitation of driest quarter
BIO18	Precipitation of warmest quarter
BIO19	Precipitation of coldest quarter

Table 3.1: Definition of the bioclimatic variables.

The monthly temperature and precipitation data was obtained from the PRISM database (Daly et al. 2002, PRISM Climate Group). The PRISM rasters have a grid cell size of 0.00833 decimal degrees and the rasters are provided in the GCS_WGS84 spatial reference system.

To calculate the bioclimatic variables we adapted the BIOVARS function from the DISMO package (Hijmans et al. 2015). The BIOVARS function from the DISMO package does not allow the user to provide a layer of the mean temperature. Instead of using the mean temperature in the calculations it uses the average of the minimum and maximum temperature. Our adaptation does use the mean temperature layers and should be slightly more accurate.

3.2.3 Normalized Difference Vegetation Index

The Normalized Difference Vegetation Index (NDVI) is an index of the amount of vegetation. It is based on measurements of the reflectance in the infra-red and the near infra-red region. The NDVI takes on values between -1 and 1. High values correspond with live green vegetation. The NDVI rasters used in this thesis came from the Global Inventory Modelling and Mapping Studies (GIMMS) and was provided by the University of Maryland Global Land Cover Facility (Pinzon, Brown, and Tucker 2005; Tucker et al. 2005). This

database contains semi-monthly rasters of the NDVI value for the period 1983-2006. The original rasters had a cell-size of 0.07266 decimal degrees. These rasters were originally

cite paper Brody / Svenning on greenes in the US

resampled to 0.04166 decimal degree rasters and cells with values corresponding to water or with no data available were set to NA. The semi-monthly rasters were then combined such that monthly rasters were obtained. To obtain an average monthly NDVI raster for each month the 24 NDVI rasters of the different years were averaged. The 12 resulting monthly NDVI rasters can then be used to calculate a minimum, maximum, and mean NDVI raster.

3.2.4 Digital elevation model

The digital elevation model (DEM) raster (*CIAT-CSI SRTM*) contains data on the elevation throughout the US. It is the raster with the highest resolution, more specifically the cell size is 0.000833 decimal degrees and the raster is provided in the GCS_WGS84 spatial reference system. The use of elevation data in SDM is somewhat contested, see e.g. Hof, Jansson, and Nilsson 2012; Oke and Thompson 2015. However, since our main purpose is to test model selection techniques adding a potentially irrelevant predictor should not matter. Furthermore, it is often suggested that other variables directly derived from DEMs, e.g. slope, are more ecological relevant (Franklin and Miller 2009). However, to keep the amount of data in this thesis handleable we will restrict ourselves to the original DEM raster.

3.2.5 Land cover

The land cover data was created from the National Land Cover Databases (NLCD) provided by the US Geological Survey. The NLCD were derived from landsat imagery of 2001 (Vogelmann et al. 2001), 2007 (Homer et al. 2007), and 2011 (Fry et al. 2011). These datasets were then transformed into rasters that utilize the Anderson level 1 classification (Anderson et al. 1976). Eight different land cover classes are used: barren, forest, ice-snow, grassland, urban, water, wetlands, agriculture. For each land cover class-year combination rasters with a cell size of 0.04166 decimal degrees were created. In order to obtain one raster for each land cover class the three corresponding rasters were averaged. The eight final rasters were then rescaled such that the values of the rasters lie within the interval $[0, 1]$. For each land cover class raster the cell values are an estimate of the percentage of the land of this class within the cell. It is interesting that the sum of these rasters equals one, hence there is a linear dependence between the variables.

3.2.6 Human Influence Index

The Human Influence Index (Wildlife Conservation Society - WCS and AU - Center for International Earth Science Information Network) raster contains integer values between 0 and 64. High values indicate a strong human influence and vice versa. The index is derived from measures of the population density, the amount of roads, the amount of light sources during night-time, etc. The raster was reprojected, see Section 3.2.7, to the GCS_WGS84 spatial reference system and has a cell size of 0.00833 decimal degrees.

3.2.7 Preprocessing of the predictor data

In order to speed up computations and facilitate general GIS operations the rasters were, if necessary, reprojected to the GCS_WGS84 spatial reference system. The extent and resolution of the rasters were set to be equal to those of the DEM layer. When necessary bilinear interpolation was used. This procedure makes sure that the cells of the rasters line up nicely. A visualization of the whole process can be found in Figure 3.1. Once all the data was preprocessed the rasters amount to over 300 GB of data.

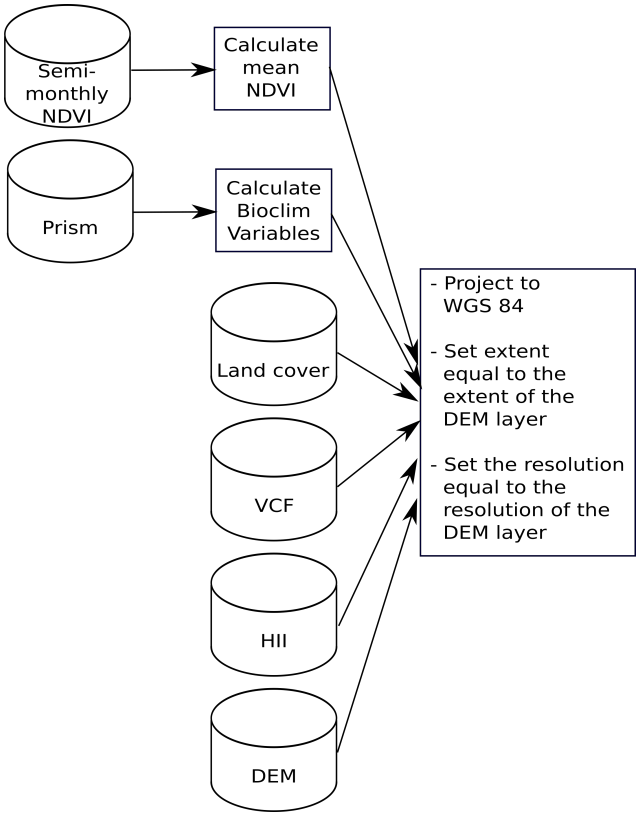

Figure 3.1: Visualization of the preprocessing of the raster data.

3.2.8 Exploratory analysis of the predictor data

It might be expected that the relationship between the variables is different in different regions of the contiguous US. To test this we defined four regions, their bounding rectangles can be found in Figure 3.2. To test our suspicions two sets of random points were generated: one with points within the contiguous United States and one that contains points within the West Coast region. For each point the corresponding values of the predictor rasters were extracted. Heat maps of the correlations of the predictor variables can be found in Figures 3.3 and 3.4. A quick inspection of these plots learns that most correlations are approximately equal in the two ranges. There are also some correlations that change quite dramatically, e.g. the correlation between the ice-snow land cover class and the NDVI indices. Even though we only report the heat maps of the correlations for the US and the West Coast region similar similar changes can be observed for the other regions.

It is interesting to note that the rank of the predictor data matrix of the random points is 32. This is rather surprising since BIO7 is a linear combination of BIO5 and BIO6 and the

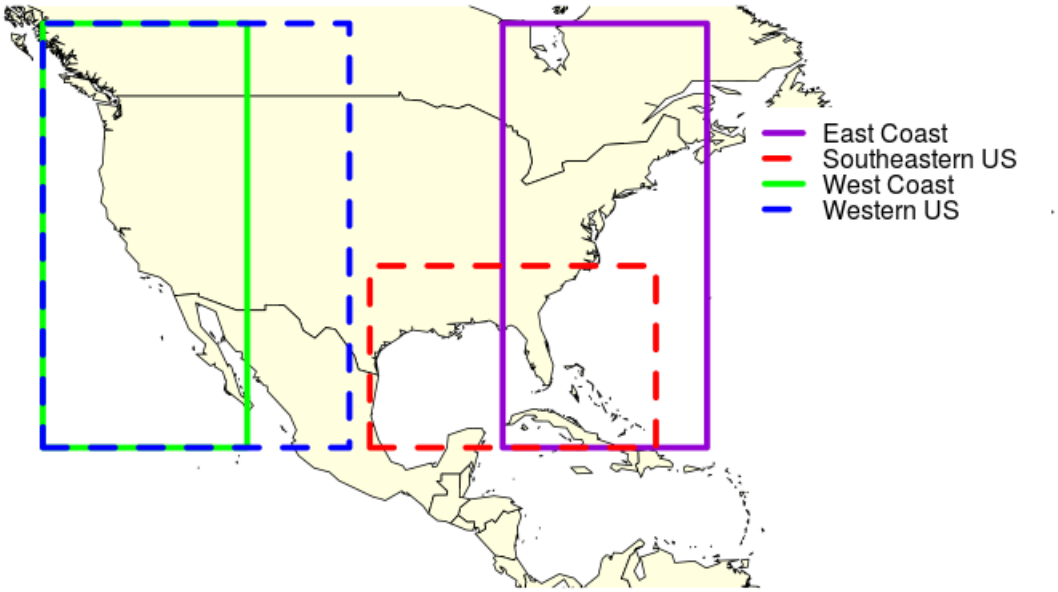

Figure 3.2: Regions of the contiguous US and their bounding rectangles.

land cover variables should sum to a constant. Hence one would expect that the rank is smaller or equal to $33 - 2 = 31$. A closer inspection leads to the conclusion that some small rounding errors in the creation of the land cover rasters "remove" the linear dependence. This "near linear dependence" also becomes clear when we look at the singular values of the scaled data matrix. The three smallest are $0.2139771, 0.000002$, and 0, for all practical purposes this means that there are two redundant variables.

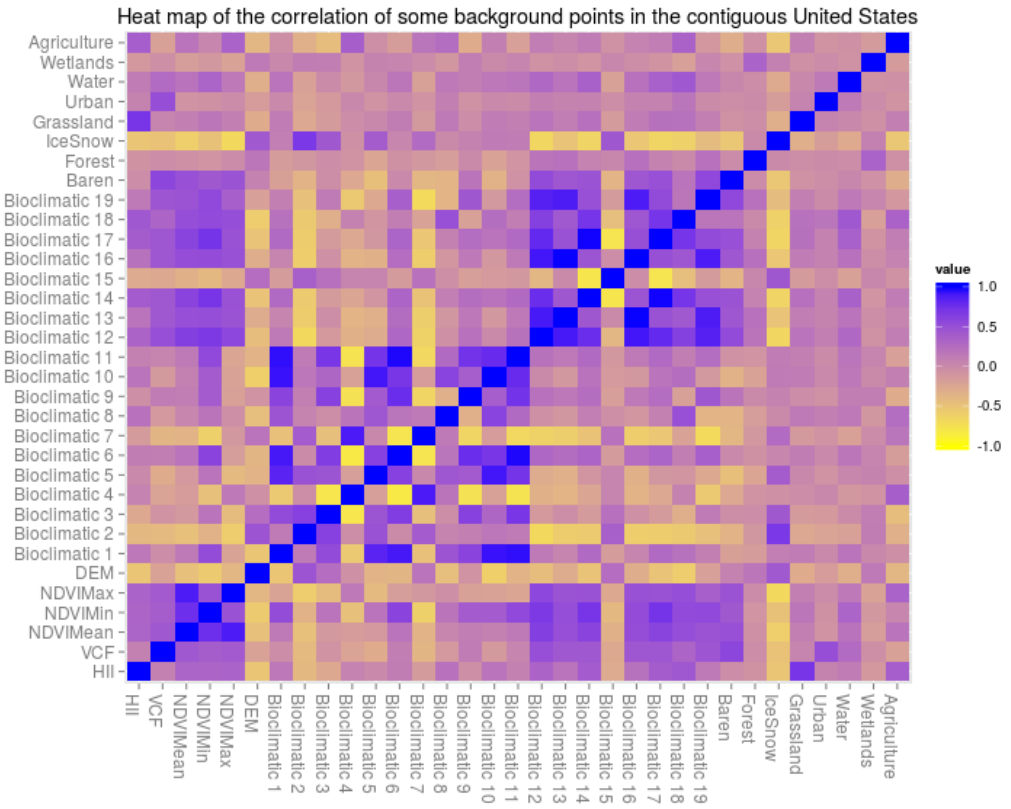

Figure 3.3: Heat map of the correlations in the contiguous US.

Figure 3.4: Heat map of the correlations in the West Coast region.

3.3 Outcome data

3.3.1 Species considered

The species that will be studied can be found in Table 3.2. These species were selected so that different regions in the US are represented. This was done because, as we saw in Section 3.2.8, the relationship between the different predictors can be different in different regions. The extent of the distributions are quite different across the selected species. For example, the copperhead snake is spread throughout a large part of the US while the Sequoia sempervirens only occurs in a small strip of land stretching from Southern California to Southern Oregon. Finally, the set of selected species consists of five plant and five animal species. Selecting both plant and animal species was done because it seems reasonable that the including fine grain predictors will lead to a larger increase in predictive performance of the classification models when the species is stationary.

Since the study areas considered are relatively large it can be expected that the bioclimatic variables will be the most important variables, see Section 2.4. Since we are specifically interested in dealing with redundant variables it should, as stated earlier, not really matter whether or not only the bioclimatic variables are important.

3.3.2 Global Biodiversity Information Facility

The presence-only data came from the Global Biodiversity Information Facility (GBIF) database. This database contains data from other smaller presence-only databases. Examples include data from citizen science projects (e.g. the iNaturalist project) or herbariums (e.g. The New York Botanical Garden Herbarium). These data sources are quite prone to errors. Citizen science data is usually provided by non-experts and misidentifications are quite likely. Even data collected by experts can be irrelevant for our purposes, for example herbarium data often includes specimens located inside botanical gardens etc. GBIF data tends to contain a lot of duplicated observations. Hence, before using the data some data-cleaning was performed. Finally, because the predictors were recorded quite recently we decided to restrict ourselves to observations obtained from the 1980s onward. To some extent this prevents the situation where the current predictor values have changed recently, e.g. due to deforestation.

3.3.3 Forest Inventory and Analysis data

The presence-absence data of the plant species was obtained from the The United States Forest Service Forest Inventory and Analysis (FIA) database. The data from this database consists of plot locations and all the tree species observed within each plot are recorded. The reported coordinates of the plots are, for privacy reasons, slightly distorted.

The sampling design that is used in the construction of this database changed in 1999 and details can be found in O'Connel et al. 2015. By 2004 the new sampling design was implemented in nearly all the states of the contiguous US. The exceptions to this are New Mexico, Oklahoma, and Wyoming for which the new design was implemented in 2005, 2008-2009, and 2011. In each state at least 10% of the plots are sampled each year. By using a time-frame of 10 years we ensured that each plot site was sampled. More

Species	Common name	US	West Coast	East Coast	Western US	Southeastern US
Aesculus glabra	Ohio buckeye	✓				
Juniperus osteosperma	Utah juniper			✓		
Quercus ilicifolia	bear oak			✓		
Salix caroliniana	coastal plain willow	✓				
Sequoia sempervirens	coast redwood		✓			
Agkistrodon contortrix Linnaeus	copperhead snake	✓				
Geomys pinetis Rafinesque	southeastern pocket gopher					✓
Pituophis catenifer catenifer	Pacific gopher snake		✓			
Sorex pacificus	Pacific shrew		✓			
Sylvilagus nuttallii	mountain cottontail				✓	

Table 3.2: The different species studied and their study extent.

Species	GBIF	FIA
Aesculus glabra	126	177
Juniperus osteosperma	230	4131
Quercus ilicifolia	98	62
Salix caroliniana	55	68
Sequoia sempervirens	717	206
Agkistrodon contortrix Linnaeus	1426	
Geomys pinetis Rafinesque	53	
Pituophis catenifer catenifer	232	
Sorex pacificus	141	
Sylvilagus nuttallii	125	

Table 3.3: Number of occurrence observations.

Region	Number of plots
US	287860
West Coast	82184
East Coast	42029
Western US	144436

Table 3.4: Number of plots within the regions.

particularly, the time-frame that was used is 2004-2014. Since the sampling in New Mexico, Oklahoma, and Wyoming started later than 2004 these states were under-sampled. States in the Eastern US tend to have a sample intensity larger than 10% and some plots were sampled multiple times. Plots that were sampled multiple times were replaced by new observations. If a plot contained the species of interest at least once the species was said to have been present, otherwise it was absent in the plot. It might be interesting to use modelling methods that allow for a sampling design correction. However, this would lead us astray and the sampling design will not be corrected for in this thesis.

Finally, all of the sampled plots are contained within forested areas. This implies that lone standing trees were not observed.

3.3.4 Data preparation

In order to build the necessary models the predictor values corresponding to the presence or absence locations are needed. For certain locations some of the rasters contained a NA value. These locations were removed before the models were constructed. The number of occurrences for each species can be found in Table 3.3. The total number of plots contained within each region can be found in Table 3.4.

Chapter 4

Classification techniques

4.1 Introduction

This chapter deals with the statistical foundations of species distribution modelling. Since it is impractical to list all the available methods we restrict ourselves to the most popular or fundamental ones. In the last 10 years a lot of research has focused on the performance of these different algorithms (e.g. Elith* et al. 2006; Segurado and Araújo 2004). The results of these studies were taken into account when the methods that are used were selected.

4.2 Presence-absence data

In this section some important and often used methods to classify binary data are reviewed. First of all, the outcome, Y_i, of observation i indicates whether a species occurs, $Y_i = 1$, or is absent, $Y_i = 0$. We denote the vector of explanatory variables as $\boldsymbol{X}$. The general form of the models used in this section is:

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\gamma}). \tag{4.1}$$

In this representation $f(\cdot;\cdot)$ is a function parametrized by $\boldsymbol{\gamma}$. The main differences between the techniques introduced below are the functional form of $f(\cdot;\cdot)$ and the loss function that is minimized.

4.2.1 Logistic regression

Perhaps the most fundamental modelling technique for binary data is logistic regression. In logistic regression the log odds ratio of the probability of an occurrence is modelled as a linear function of the covariates. Hence, the model can be depicted as

$$\log\left(\frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{P(Y = 0|\boldsymbol{X} = \boldsymbol{x})}\right) = \gamma_0 + \boldsymbol{x}^t\boldsymbol{\gamma}.$$

It is easy to show that this model can be written in the form used in Equation 4.1. More specifically, if we define $\text{expit}(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ we get

$$f(\boldsymbol{x}; \boldsymbol{\gamma}) = \text{expit}(\gamma_0 + \mathbf{x}^t\boldsymbol{\gamma}).$$

Usually the coefficients of a logistic regression model are obtained by using maximum likelihood estimation (MLE). Obtaining the MLE $\widehat{\boldsymbol{\gamma}}$ corresponds with solving the following maximization problem:

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \{y_i \log f(\boldsymbol{x}_i; \boldsymbol{\gamma}) + (1 - y_i) \log f(\boldsymbol{x}_i; \boldsymbol{\gamma})\}.$$

When one multiplies this log-likelihood function by minus one we get a loss function that is often called the cross-entropy. For more information about logistic regression and numerical optimization techniques for obtaining the MLE we refer to Agresti 2013; McCullagh and Nelder 1999.

The main advantages of logistic regression models are that they are relatively simple to implement, interpret, etc. This simplicity is also its greatest disadvantage. In particular, when modelling the distribution of a species there is often no a priori knowledge of the shape of the response curves.

4.2.2 Generalized additive models

In standard logistic regression a linear systematic component is used. It is easy to extend logistic regression models to include non-linear systematic components. However, the functional form of the log odds ratio might not be known by the researcher and hence a non-parametric (or semi-parametric) modelling technique can be useful. When the distribution of the outcome belongs to the exponential family generalized additive models (GAMs) are one possible class of non-parametric (or semi-parametric) models. In the case of a Bernoulli distribution the resulting GAM is sometimes called an additive logistic regression model and has the form:

$$\log\left(\frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{P(Y = 0|\boldsymbol{X} = \boldsymbol{x})}\right) = \gamma_0 + f_1(x_1) + \cdots + f_p(x_p). \tag{4.2}$$

In this representation the $f_k(\cdot)$'s are certain smooth functions. Again one can write this model in the form of Equation 4.1:

$$f(\boldsymbol{x}; \boldsymbol{\gamma}) = \gamma_0 + f_1(x_1) + \cdots + f_p(x_p)$$

There is a multitude of popular ways to specify the $f_k(\cdot)$'s (Hastie and Tibshirani 1990; Wood 2006). We will follow Wood 2006; Wood and Augustin 2002 and focus on using cubic smoothing splines to represent the $f_k(\cdot)$'s.

GAMs are most often fitted by using MLE. In order to restrict the "wiggliness" of the smoothing functions in model 4.2 one can add a penalization term to the likelihood function. An example of such a "wiggliness" penalty is:

$$\sum_{j=1}^{p} \lambda_j \int_{x_{j(1)}}^{x_{j(n)}} \{f_j^{(2)}(x)\}^2 dx.$$

In this penalization term the $x_{j(1)}$ (resp. $x_{j(n)}$) is the smallest (resp. largest) value of the j'th covariate. Furthermore, it can be shown that, in a class of sufficiently smooth functions, the minimizer of

$$-\sum_{i=1}^{N} \{y_i \log f(\boldsymbol{x}_i; \boldsymbol{\gamma}) + (1 - y_i) \log f(\boldsymbol{x}_i; \boldsymbol{\gamma})\} + \sum_{j=1}^{p} \lambda_j \int_{x_{j(1)}}^{x_{j(n)}} \{f_j^{(2)}(x)\}^2 dx$$

is a natural cubic spline with knots at the n covariate values. The two limiting cases, $\lambda = 0$ and $\lambda = \infty$, are interesting. If $\lambda = 0$ an interpolating spline is optimal, while when $\lambda \to \infty$ the solution converges to the linear logistic regression solution. Finally, fitting GAMS with natural cubic splines as smoothing functions is, compared to most other smoothers used in GAMs, computationally efficient.

In practice we have to select the penalization parameters, this is usually done by using cross-validation. Because the fitting procedures for GAMs are computationally intensive usually the closely related generalized cross-validation (GCV) is used (Wood and Augustin 2002). Finally, since $\lambda = \infty$ still allows a first order fit, it can be interesting to perform additional model selection. Although variable selection is the topic of Chapter 5 we mention that it is possible to introduce an extra penalization term that leads to an automatic variable selection technique (Marra and Wood 2011), i.e. might remove a predictor from the model. This approach is implemented in the MGCV package (Wood 2015).

Finally, up till now only univariate splines have been considered, if interaction terms are to be included one can use e.g. thin plate splines. However, using multidimensional splines usually results in a computationally intensive fitting procedure. Finally, for a more applied review of the use of GAMs in species distribution modelling we refer to Guisan, Edwards Jr, and Hastie 2002.

4.2.3 Artificial neural networks

Artificial neural networks (ANNs) are a non-linear modelling technique. We refer to Bishop 1995 for an introduction to the general methodology and some of the technical details. The terminology used in the ANN literature is slightly different than in the standard statistical literature. More particularly, the explanatory variables are usually called the input features. Furthermore, an ANN consists of so-called "layers" of "neurons". The first layer is called the input layer and consists of one neuron for each variable. In each successive layer the output of the corresponding neurons is the result of applying an activation function, $g(\cdot)$, to a linear combination of the values from the previous layer. The coefficients of these linear combinations are called the weights of the ANN. These weights are the parameters one can tune. The process of feeding the values of the previous layer into the next is repeated up to the last layer which is called the output layer. The layers that are neither the input nor output layer are called hidden layers. A graphical representation of an ANN with one hidden layer can be found in Figure 4.1.

From here on we will denote the vector of all the weights as $\boldsymbol{\gamma}$. As usual, the estimated weights are obtained by minimizing a loss function. When ANNs are applied to classification problems often either the squared loss, $L(\boldsymbol{y}; \boldsymbol{\gamma}) = \sum_i (y_i - f(\boldsymbol{x}_i; \boldsymbol{\gamma}))^2$ or the cross-entropy, $L(\boldsymbol{y}; \boldsymbol{\gamma}) = -\sum_i [y_i \log\{f(\boldsymbol{x}_i; \boldsymbol{\gamma})\} + (1 - y_i) \log\{f(\boldsymbol{x}_i; \boldsymbol{\gamma})\}]$ is used. We will use the cross-entropy since it is closely related to the likelihood of a Bernoulli distribution and hence more in line with the other techniques. To minimize the loss criterion, often backpropagation (Rumelhart, Hinton, and Williams 1986) is used in combination with a numerical optimization algorithm, e.g. steepest descent.

It is easy to see that logistic regression can be seen as an ANN with: no hidden layers, the expit function as the activation function of the output layer, and the cross-entropy loss.

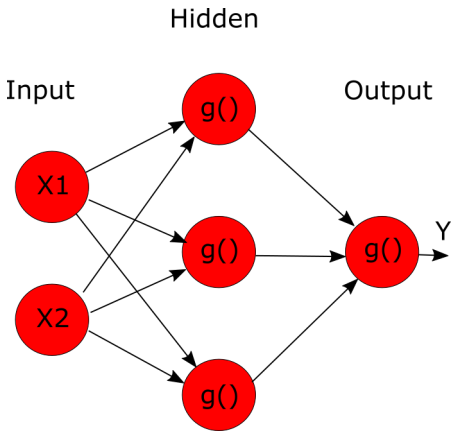

Figure 4.1: Visualization of a feed-forward neural network with one hidden layer.

The biggest strength of ANNs is that, under some regularity constraints, they can approximate any continuous function arbitrarily well (Hornik, Stinchcombe, and White 1989). Some disadvantages include:

- Selecting an optimal number of layers and neurons is far from trivial.
- The loss-function often has multiple local minima.
- Different numerical optimization methods often lead to different solutions.
- Fitting large neural network architectures can be computationally infeasible.
- The backpropagation algorithm cannot be used in combination with non-differentiable penalty functions, e.g. the lasso penalty (see Section 5.3.2).
- The fitted parameters can be sensitive to the initial weights.
- The obtained model seems like a "black-box", i.e. there is often no easy way to interpret the parameters and the effect of different predictors.

Some of these disadvantages can be, partially, overcome by using e.g. weight decay, averaging networks, early stopping, pruning, ... Weight decay is also referred to as L_2 regularization and is described in Section 5.3.1. Averaging networks boils down to using different sets of initial values, fitting the same network structure for each set, and then combining these into one model (Ripley 2009). The CARET package (Kuhn et al. 2015) provides an implementation of averaged neural networks based upon the NNET package (Ripley and Venables 2016).

4.2.4 Tree based methods

In this section some tree based methods are introduced. The section starts with a short explanation of decision trees. Afterwards boosting is introduced as a method to deal with some of the shortcomings of decision trees.

Decision trees

Tree based methods are a class of algorithms that partition the input space into rectangular regions. The same predicted value is then assigned to all observations within a certain region. In the context of SDMs we can interpret this as partitioning the environmental space into rectangles. Each of these rectangles is then labelled as being part of the niche or not. To obtain such rectangles we start by splitting the input space into two regions along one variable. The variable and the value that is used to obtain this split is selected in such a way that the loss function of interest is minimized. In the following steps either the algorithm stops if some stopping criterion is met or the obtained sub-regions are split into smaller sub-regions. After the tree is grown usually a winner takes all approach is applied to obtain the label for each region. Thus, if a region contains mainly presences its predicted class will be the presence class, otherwise it will be labelled as a region containing absences. Finally, the number of nodes J of a tree is defined as the number of splits $+1$. A visualization of a partitioning obtained bys using a classification tree is given in Figure 4.2.

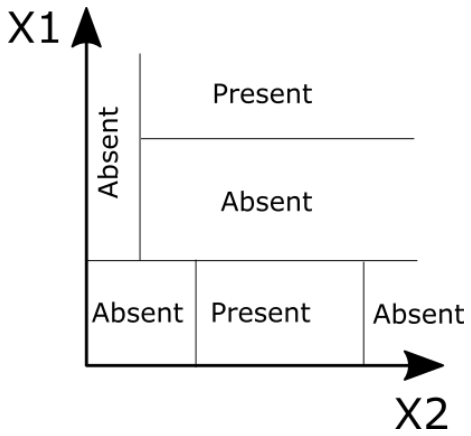

Figure 4.2: Visualization of a classification tree in a two dimensional input space.

Unless a stopping criteria is specified this approach would lead to over-fitting. Hence, the algorithm is usually stopped when no new splits can be found that decrease the loss by some pre-specified amount. Another possible stopping criterion is to stop the algorithm once a certain number of splits is reached. As usually we opt for using the cross-entropy as the loss function. Finally, we note that there are many variations to the algorithm sketched above.

The most important advantages of decision trees include:

- Complex interaction effects can easily be modelled.
- Small decision trees are easily visualized.
- Decision trees usually perform relatively well even when no variable selection was applied.
- The idea underlying decision trees is quite intuitive.
- Decision trees are invariant under monotonic transformations of the predictors.

Some of the disadvantages of using decision trees include:

- Decision trees usually have a large variance
- Categorical variables with a lot of classes can lead to computational problems.

Boosting

Boosting can be used to combat the large variance of decision trees. The overview of boosting presented in this section is largely based on Elith, Leathwick, and Hastie 2008; Friedman, Hastie, and Tibshirani 2000. Elith, Leathwick, and Hastie 2008 gave an applied working guide to boosted regression trees while Friedman, Hastie, and Tibshirani 2000 gave a theoretical explanation of boosted regression trees.

An ensemble of different classifiers is sometimes useful to obtain improved classifiers. This is what is done in boosted classification trees. Boosting can be described as creating a sequence of models such that the i'th model focusses on correctly classifying observations that were misclassified by the $i-1$ previous models. The corresponding algorithm then takes the following form:

1. Construct a classifier.
2. Classify the observations.
3. Assign large weights to wrongly classified observation and vice versa for correctly classified observations.
4. Fit a classifier on the weighted data-set.
5. Create a new classifier by adding the the new and old classifiers, usually the new classifier is shrunken by multiplying it with a regularization parameter.
6. Stop if some stopping criteria is met, otherwise use the new classifier obtained in step 5. and repeat from step 2. onwards.

In order to avoid over-fitting one can use a new subsample of the full data-set in each iteration of the algorithm. When combining boosting with classification trees it is recommended that the subsamples have a sample size in between 0.5 and 0.75 times the size of the full dataset (Elith, Leathwick, and Hastie 2008).

Another popular ensemble method that is used in combination with decision trees is the Random Forest (RF) method. In most cases RFs perform slightly worse than boosted classification trees (Hastie, Tibshirani, and Friedman 2009) and we will not consider them in this thesis.

A disadvantage of boosted classification trees is that they are not as easily visualized / interpreted as normal classification trees. Secondly, there are quite a few tuning parameters, namely the depth of the trees J, the regularization term, and the number of trees. However, Hastie, Tibshirani, and Friedman 2009 note that using $J \in \{4, \ldots, 8\}$ usually leads to the best performing models. Additionally they observed that the specific value of J in this set has little effect on the performance of the classifier. Hence, cross validation will be used to obtain optimal values for the shrinkage parameter and the number of trees.

Finally, although we described boosting for classification trees the same idea can be readily applied to other classification algorithms. Furthermore, boosted classification trees seem to include a form of internal variable selection, i.e. the performance of this method usually doesn't degrade a lot when irrelevant predictors are present. We will use the implementation provided by the GBM package (Ridgeway 2015).

4.3 Presence-only data

Instead of having access to presence-absence data it happens quite often that only presence data is available. Since the previously described classification techniques need binary data they can not immediately be used with presence-only data. In Section 4.3.1 the inhomogeneous Poisson process (IPP) is introduced. Presence-background classification is introduced in Section 4.3.2. Finally, in Section 4.3.3 Maximum Entropy (MaxEnt; Phillips, Anderson, and Schapire 2006; Phillips and Dudík 2008) is concisely described.

Although we will not consider it we note that there is another interesting way to use regression models in combination with presence-only data. Ward et al. 2009 used the EM algorithm (Dempster, Laird, and Rubin 1977) in combination with regression models. Although this leads to an elegant and rigorously motivated method of fitting regression models, the prevalence of the species needs to be known, or estimable, which is nearly never the case.

4.3.1 Poisson point processes

First of all, all the measure theoretic machinery involved in point processes is blatantly ignored in this section. We denote the study area of interest by S. Usually S corresponds to an area preserving projection of the earth and therefore our focus is on $S \subseteq \mathbb{R}^2$. A point process is a random variable $\boldsymbol{X}$ that consists of points $u_i \in S$, hence $\boldsymbol{X} = (u_1, \ldots, u_n)$. In the presence-only scenario the $\boldsymbol{X}$ variable equals the locations where an occurrence was reported. One popular point process is the inhomogeneous Poisson point process. Before giving a non-rigorous definition of the IPP we define the intensity of a point process as a function $\lambda(\cdot) : S \to \mathbb{R}_+$ for which $\int_B \lambda(u)du < \infty$, $\forall$ bounded $B \subset S$. The random variable $\boldsymbol{X}$ is said be an IPP if $\forall$ bounded $B \subset S$:

1. The number of observed points within B, $N(B) = \#(X \cap B)$, follows a Poisson distribution with mean $\Lambda(B) = \int_B \lambda(u)du$.

2. Given the number of observed points within B, $N(B)$, the elements of X within B are i.i.d. distributed with density $\frac{\lambda(u)}{\Lambda(B)}$.

The first part of the definition implies that the the expected number of observations in a small area surrounding a point is approximately equal to the intensity, assuming it is a continuous function, multiplied by size of the area.

In order to obtain a parametric model of an IPP the intensity is usually modelled as a function of the covariates $\boldsymbol{z}(u)$ observed at the location u. A popular model is the log linear model:

$$\lambda(u) = \exp(\gamma_0 + \boldsymbol{\gamma}^t \boldsymbol{z}(u)).$$

It is interesting to note that the intercept term, γ_0, only affects the expected number of points but not the configuration of the points in S. When we use C to denote a rest term that is independent of the parameters, the likelihood function becomes:

$$\begin{aligned} l(\boldsymbol{x};\boldsymbol{\gamma}) &= \log\left(\prod_{i=1}^{n}\frac{\lambda(u_i)}{\Lambda(S)}\frac{\Lambda(S)^n\exp(-\Lambda(S))}{n!}\right) \\ &= \sum_{i=1}^{n}\left\{\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{z}(u_i)\right\} - \int_S \exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{z}(u))du + C. \end{aligned}$$

The difficulty of using MLE in combination with IPP models is approximating the integral. For an overview of different methods to obtain the MLEs and a more complete treatment of spatial point processes we refer to Møller and Waagepetersen 2007.

4.3.2 Classification with pseudo-absences

One of the older techniques to deal with presence-only data consists of generating so-called pseudo-absence or background points. The original motivation of this technique is that by uniformly sampling n_0 points in the geographical space S one obtains a representation of the available habitat in the study region (Pearce and Boyce 2006). The available habitat is then contrasted with the locations where a presence was observed by using standard classification methods. In practice one has to decide on the number of pseudo-absences that should be used and whether or not the pseudo-absences should be weighted (Barbet-Massin et al. 2012).

Another, more rigorous, way to justify the pseudo-absence method is by using the IPP model. Warton and Shepherd 2010 were the first to note the connections between fitting an IPP model and fitting logistic regression models with pseudo-absences. Fithian and Hastie 2013 further refined these equivalences and also considered the situation when the model is misspecified. We follow the derivation given by Fithian and Hastie 2013.

First of all, we condition on the number of presence points, n_1, and the number of background points, n_0, and view this as a case-control sampling scheme. It is clear that the probability of outcome Y_i being a presence point equals $P(Y_i=1)=\frac{n_1}{n_0+n_1}$ and hence $P(Y_i=0)=\frac{n_0}{n_0+n_1}$. Now suppose that the presence points are generated by a IPP with intensity $\lambda_1(u)=\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{z}(u))$. Since the background points are generated uniformly over S the intensity of this process is $\lambda_0(u)\propto 1$. Conditional on n_1 and n_0 it is now easy to derive the expression of a logistic model:

$$\begin{aligned} P(Y_i=1|U=u) &= \frac{f(u|Y_i=1)P(Y_i=1)}{P(Y_i=1)f(u|Y_i=1)+P(Y_i=0)f(u|Y_i=0)} \\ &= \frac{\frac{\lambda_1(u)}{\Lambda_1(S)}\frac{n_1}{n_0+n_1}}{\frac{\lambda_1(u)}{\Lambda_1(S)}\frac{n_1}{n_0+n_1}+\frac{\lambda_0(u)}{\Lambda_0(S)}\frac{n_0}{n_0+n_1}} = \frac{\frac{\lambda_1(u)n_1\int_S du}{\Lambda_1(S)\lambda_0(u)n_0}}{\frac{\lambda_1(u)n_1\int_S du}{\Lambda_1(S)n_0}+1} \\ &= \frac{\exp\left(\gamma_0+\log\left(\frac{n_1\int_S du}{n_0\Lambda_1(S)}\right)+\boldsymbol{\gamma}^t\boldsymbol{z}(u)\right)}{\exp\left(\gamma_0+\log\left(\frac{n_1\int_S du}{n_0\Lambda_1(S)}\right)+\boldsymbol{\gamma}^t\boldsymbol{z}(u)\right)+1} \end{aligned} \tag{4.3}$$

This derivation implies that, with the exception of the intercept, the coefficients of the IPP model can be estimated by using the standard fitting procedures for logistic regression.

Furthermore, it shows that combining logistic regression together with back-ground data is quite a natural approach to the presence-only problem. The fact that the intercept of the IPP model cannot be obtained is not problematic. As we saw in Section 4.3.1 the intercept does not affect the configuration of the observed points within S. Instead it basically rescales the expected number of observed presences. Since the number of observed presences is partially determined by processes of which we have no data, e.g. the sampling intensity, the intercept is usually not of interest.

4.3.3 Maximum Entropy modelling

Perhaps the most popular method to create models from presence-only data is Maximum Entropy modelling. Phillips, Anderson, and Schapire 2006 considered a gridded study area $\mathcal{X}$. An occurrence of a species then occurs in a cell based manner, hence instead of an exact location in S it is only known that the species was present inside the corresponding cell in $\mathcal{X}$. They then went on to find a distribution $\pi(x) : \mathcal{X} \to \mathbb{R}$ which minimizes the entropy:

$$-\sum_{x \in \mathcal{X}} \pi(x) \log\{\pi(x)\},$$

under the constraint that the observed mean of the features, $\boldsymbol{z}(x) \in \mathbb{R}^p$, is "close" to the expected mean. If we denote the set of cells for which an occurrence was observed by $\mathcal{X}_1$ the constraint is:

$$\left| \sum_{x_i \in \mathcal{X}_1} \boldsymbol{z}(x_i) - \sum_{x \in \mathcal{X}} \pi(x) \boldsymbol{z}(x) \right| < \boldsymbol{\lambda}, \quad \boldsymbol{\lambda} \in \mathbb{R}^p_+,$$

where the inequality is applied component wise. It can be shown that the solution to this problem has the following form:

$$\pi(x) \propto \exp(\boldsymbol{\gamma}^t \boldsymbol{z}(x)), \quad \boldsymbol{\gamma} \in \mathbb{R}^p.$$

It is interesting to note that this is exactly the form of a log linear IPP model. Renner and Warton 2013 showed that when the grid size becomes small the MaxEnt solution converges towards the penalized IPP solution. The standard implementation of MaxEnt uses as features a combination of products between covariates, hinge features, step functions, quadratic terms and linear terms (Phillips and Dudík 2008). Hence, in the end MaxEnt modelling is equivalent with a penalized version of presence-background logistic regression in an extended covariate space.

4.4 Taking the scale hierarchy into account

In Section 2.4 the influence of spatial scale was discussed. There have been some attempts to incorporate this hierarchical structure into the classification techniques. E.g. Pearson, Dawson, and Liu 2004 combined two models, one for coarse scale processes and one that introduces fine grain variables. More specifically their approach was as follows:

1. An initial model is obtained by solely using climate variables as predictors.
2. The predicted values are saved as a new variable.

3. The predicted values and remotely sensed variables are combined and used as the input for a second model.

4. The predicted values of the second model are the final predictions.

Graphically this can be depicted as in Figure 4.3.

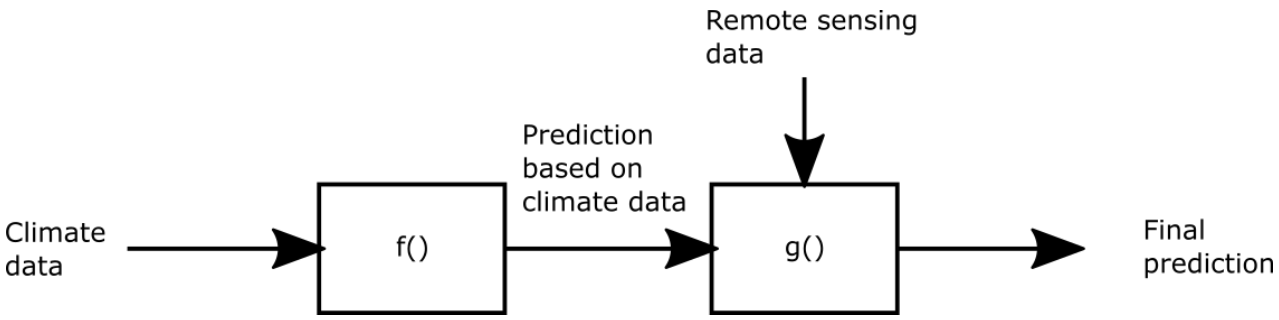

Figure 4.3: Visualization of the hierarchical model.

This hierarchical model has since been applied in combination with e.g. stacked SDMs (Cord et al. 2014). However, from a statistical point of view this approach is not well motivated. The main drawback of this approach is that interaction effects between the climate and remotely sensed variables cannot be taken into account. Furthermore, since there is usually quite some correlation between the climate and remotely sensed variables model selection is hampered. For example, if a remotely sensed variable is fundamental in determining the niche and highly correlated with a climate variable that is not as relevant for defining the niche, then the climate variable will usually end up in the model instead of the remotely sensed variable. Because of these drawbacks this approach will not be investigated.

Chapter 5

Reducing the number of explanatory variables

5.1 Introduction

The explanatory variables used in SDMs are often correlated. Highly correlated variables can be an indication of redundant information in the data-set. Moreover, irrelevant predictors might be included into the model since there is usually no knowledge about which variables make up the niche. It is well known that this can lead to over-fitting and unstable predictions. In this chapter we describe some methods to deal with large sets of correlated predictors. More specifically, in Section 5.2 we introduce methods that transform the input space. In Section 5.3 techniques that penalize "large" models are introduced. Finally, Section 5.4 deals with step-wise variable selection methods.

For a review of methods to deal with correlated covariates in ecology we refer to Dormann et al. 2013. Finally, these automatic selection procedures are of course not meant to replace a well founded motivation of why certain predictors should be selected. A discussion of how the available data, scale of the predictors, etc. should influence the decision of using a complex or a simple model can be found in Merow et al. 2014.

5.2 Dimensionality reduction

Dimensionality reduction techniques can be used to obtain a new, often lower-dimensional, representation of important structures in the input space. This can be particularly useful when two or more explanatory variables are proxies of one underlying latent variable. For example, there are multiple indices that indirectly measure the amount of vegetation. A combination of these indices might be a better indicator of the amount of vegetation than the individual indices. The dimensionality reduction techniques that are used in this thesis use only the input space and ignore the outcome values. It should however be noted that there are other dimensionality reduction techniques that do take into account the relationship between input and output variables, e.g. partial least squares (see e.g. Marx 1996).

5.2.1 Principal component analysis

Often principal component analysis (PCA) is introduced as a method which constructs uncorrelated linear combinations of the variables, these new variables are called principal components. However, for our purposes it might be more interesting to view PCA as a way to find a low dimensional affine subspace such that when the original data is projected onto this subspace the "information loss" is minimal. One possible characterization of PCA is that a set of K orthonormal vectors $\boldsymbol{u}_i$ and an offset vector $\boldsymbol{b}$ are constructed so that

$$\sum_{i=1}^{N}\sum_{j=1}^{K} ||\boldsymbol{x}_i - \boldsymbol{x}_i^t \boldsymbol{u}_i \boldsymbol{u}_i - \boldsymbol{b}||_2^2$$

is minimized. It can be shown that for a certain K this sum is minimized when the $\boldsymbol{u}_i$ are the eigenvectors corresponding with the K largest eigenvalues of the covariance matrix. A prototype scenario is shown in Figure 5.1.

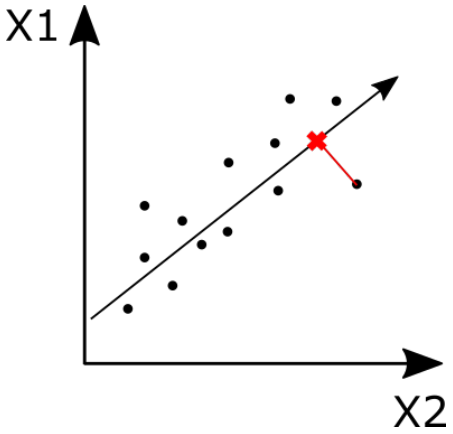

Figure 5.1: Visualization of the a typical scenario where PCA is useful.

Once the principal components have been obtained they can be used as the explanatory variables in one of the models of Chapter 4. Since the new variables are uncorrelated and some irrelevant noise in the predictors might be removed the resultant models often exhibit less variance.

There are multiple criteria based upon which the number of principal components, K, can be chosen. One possibility is to plot the so-called "explained" variance versus the number of components and look for either a kink or a point where a certain percentage of variance is explained. If PCA is combined with a classification method it is usually more appropriate to use cross-validation to select the number of principal components. We will use the cross-validation approach.

It should be clear that the main disadvantage of PCA is that it only allows for linear representations of the data. Hence, PCA will often be useless if the observations are scattered around a non-linear manifold. For example in Figure 5.2 there is a clear underlying space that is one dimensional. However, using the first principal component of this fictional data-set would lead to as big an "information" loss as using just one of the original axes. For the sake of completeness we mention that there are extensions of PCA that use non-linear manifolds instead of linear ones, e.g. principal curves and surfaces

(Hastie and Stuetzle 1989).

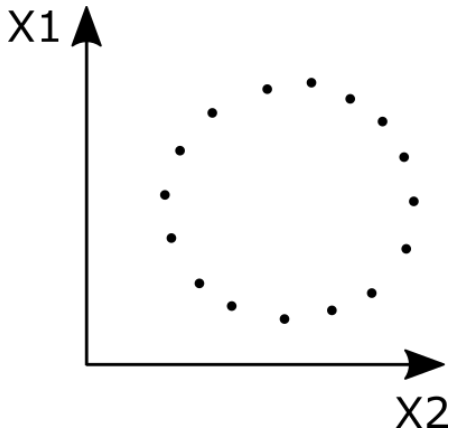

Figure 5.2: Visualization of the a scenario where PCA is useless.

5.2.2 Kernel principal component analysis

A popular and computationally efficient non-linear dimensionality reduction technique is kernel PCA (Schölkopf, Smola, and Müller 1997). In kernel PCA the elements of the input space $\mathcal{X}$, in our case we have $\mathcal{X} = \mathbb{R}^p$, are mapped to a Hilbert space F. We will denote the map as:

$$\boldsymbol{\phi}(\cdot) : \mathcal{X} \to F.$$

In this new vector space a PCA is conducted and new coordinates are obtained. A visual presentation of these steps is given in Figure 5.3.

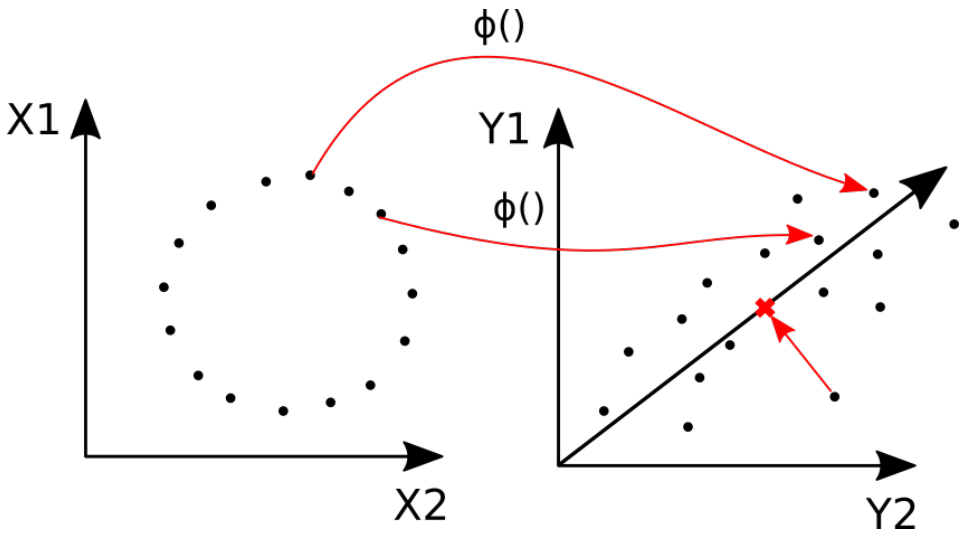

Figure 5.3: Visualization of kernel PCA.

One attractive property of kernel PCA is that we do not need to actually compute $\phi(\boldsymbol{x})$. More specifically we only need to compute the so-called kernel matrix $\mathbb{K}$

$$\mathbb{K}_{ij} = \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

In general it can be shown that, under some regularity conditions, for any kernel function

$$k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$$

there exists a corresponding $\boldsymbol{\phi}(\cdot)$ and Hilbert space F. Hence usually one specifies the kernel function instead of the map $\boldsymbol{\phi}(\cdot)$.

Popular kernel functions are:

- The polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^t\boldsymbol{y} + c)^d$.
- The radial basis kernel (RBF kernel) $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{\sigma}\right)$, with $\sigma > 0$.
- The hyperbolic tangent kernel, $k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\sigma\boldsymbol{x}^t\boldsymbol{y} + c)$.

In these representations the σ, resp. c, parameter can be seen as a scale, resp. offset, parameter. The value of these parameters is in general selected by using cross validation. We will focus on the RBF kernel, this kernel is usually used when no prior information is available (Karatzoglou et al. 2004).

A disadvantage of kernel PCA is that it is not always clear which kernel should be used. For our goals, combining kernel PCA and a classification method, we can use cross-validation and treat the type of kernel as a tuning parameter. Furthermore, it is not always particularly clear what the new features are supposed to represent (we might not even know in which Hilbert space we are working). An R implementation of kernel pca is provided as part of the KERNLAB package (Karatzoglou et al. 2004).

5.2.3 Presence versus background data

In this section we focus on PCA but all the comments readily apply to kernel PCA.

When PCA is combined with classification techniques one usually performs the PCA on the predictor values of all the included cases. However, in presence-only models the background points are randomly generated and the number of these points can be chosen. It should be clear that when the number of background points approaches infinity the covariance matrix used in the PCA converges to the covariance matrix of the explanatory data within the study extent. Unless the covariance matrix of the explanatory variables within the cells where the species is present equals the covariance matrix over the whole study extent, performing PCA on only the values corresponding with the presence points will lead to different results. For these reasons it was decided to test whether or not the effect of performing PCA on the presence, background, or all points has a significant effect on the performance of the final SDMs.

5.3 Regularization

When one uses regression methods, e.g. logistic regression or ANNs, in combination with a large set of correlated predictors the obtained coefficients are often excessively large and unstable. To combat this large coefficients can be penalized such that the solution consists of shrunken coefficients. To do this the standard minimization problem

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} L(\boldsymbol{y}, \boldsymbol{\gamma})$$

is adjusted to

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} L(\boldsymbol{y}, \boldsymbol{\gamma}) + J(\boldsymbol{\gamma}, \boldsymbol{\lambda}). \tag{5.1}$$

In this representation the function $J(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ is usually a monotonically increasing function in $\boldsymbol{\gamma}$. Furthermore, the $\boldsymbol{\lambda}$ parameters are used to control "the amount of regularization" and are often selected by using cross-validation. Finally, the described regularization methods can be used in conjunction with logistic regression by using the GLMNET package (Friedman et al. 2015).

5.3.1 Ridge regression / L_2 regularization

Ridge regression (also called L_2 regularization) is obtained when the Euclidean norm of the coefficients is used as the penalization function:

$$J(\boldsymbol{\gamma}, \lambda) = \lambda ||\boldsymbol{\gamma}||_2^2.$$

We immediately see that small λ's correspond to a small amount of regularization and the non-penalized solution is obtained when $\lambda = 0$. Furthermore, when $\lambda \gg$ the coefficients are shrunken to zero. A typical evolution of the coefficients in function of the regularization parameter can be found in Figure 5.4.

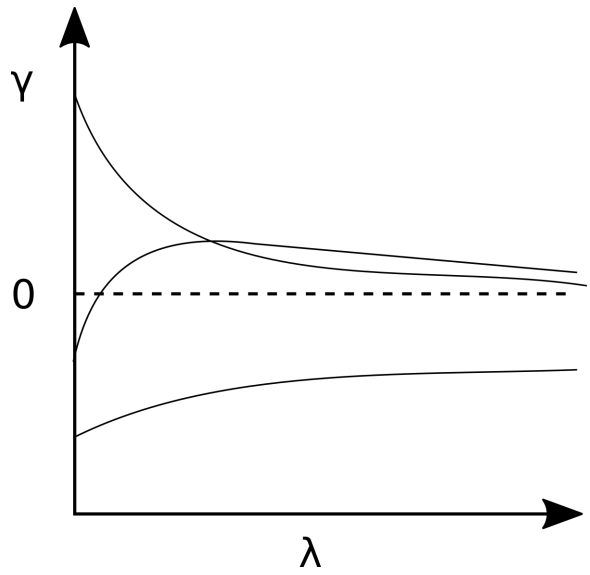

Figure 5.4: A stereotypical evolution of the regression parameters in combination with the ridge parameter.

It is clear that rescaling the covariates will usually lead to different penalized coefficients. It is therefore recommended to standardize the covariates before applying ridge regression.

Advantages of using an L_2 penalty include:

- The penalty is differentiable and hence compatible with e.g. backpropagation.
- For most regression problems it is easy to adopt the standard algorithms to include the L_2 penalty.

- It is usually no problem if the number of variables is larger than the number of observations.

The biggest disadvantage is that all the predictors are kept in the model, i.e. usually no coefficients are equal to zero.

Finally, we note that L_2 regularization is also called weight decay when it is used in combination with neural networks.

5.3.2 Lasso / L_1 regularization

Another popular option was proposed by Tibshirani 1996. He suggested to use the absolute norm (also called the L_1 norm or Manhattan distance) as penalty function, or thus the minimization problem becomes:

$$J(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \lambda ||\boldsymbol{\gamma}||_1 = \lambda \sum_{i=1}^{p} |\gamma_j|.$$

The L_1 penalty term is also called the least absolute shrinkage and selection operator (lasso). Unlike L_2 regularization, the lasso solution will usually contain some coefficients that are equal to zero. This implies that, in addition to the parameter shrinkage, the lasso performs some form of automatic variable selection. A typical parameter trace can be found in Figure 5.5.

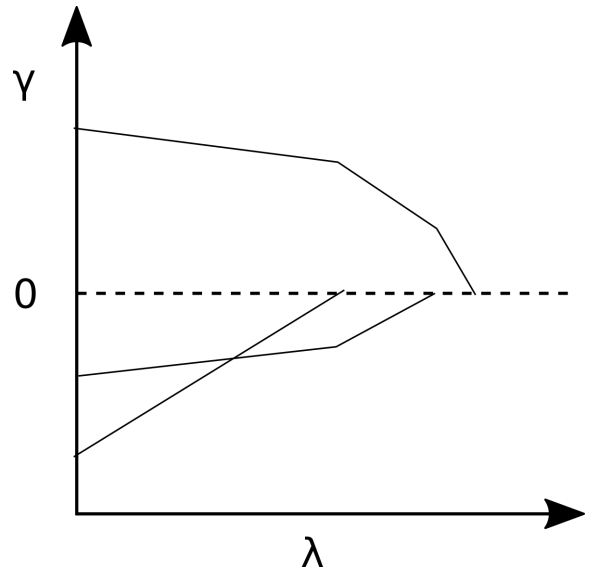

Figure 5.5: Visualization of the typical evolution of the regression parameters in combination with the lasso parameter.

Usually the shrinkage parameter is selected by using cross-validation. The main advantage is that the lasso performs both shrinkage and parameter selection at once. Disadvantages include:

- The lasso penalty is not differentiable and hence one cannot use algorithms like backpropagation.

- At most n variables are included in the model.
- If there is a group of highly correlated variables usually only one will be selected. Sometimes the variable that is selected is rather arbitrary.

5.3.3 glmnet implementation

When the cross-validation implementations of the lasso or ridge logistic regression from GLMNET package are used, the selected λ's do not correspond with 5.1. Instead the largest λ within one standard error of the optimal criterion is selected. The reasoning behind this is that there is no convincing evidence that the selected λ is better than the optimal λ, however using a larger shrinkage parameter usually, in the case of the lasso, leads to a sparser model.

5.4 Subset selection methods

Subset selection methods are a set of older methods to reduce the number of explanatory variables. Although subset selection methods are often criticised for ignoring problems with bias, multiple testing, etc. (Whittingham et al. 2006) they are still popular. In this section three different subset selection techniques are discussed.

5.4.1 Best subset selection

The most elementary technique in this set of methods is best-subset selection. Best subset selection consists of fitting a model for each combination of predictors and then selecting the best model from these. What constitutes the best model depends on the goals of the study but often used selection criteria include: AIC, missclassifcation error, p-values, etc. The biggest limitation of best-subset selection is that when the number of predictors increases there is an exponential increase in computational complexity. In particular, if there are p potential terms to be included we need to fit 2^p models. Since we have 33 variables this method is infeasible for our purposes.

5.4.2 Stepwise subset selection

A second subset selection method is backward-stepwise selection. This method starts with the model containing all p predictors. In the second step of the algorithm we try removing each predictor from the full model and select the most optimal model from the models with $p-1$ covariates. In the third step we remove each predictor from the model obtained in the second step, fit a model with $p-2$ predictors, and select the best model from these. This process is repeated until there is no improvement possible. One of the most important disadvantages of this method is that it is quite variable. Furthermore, for some algorithms (e.g. logistic regression) we need to have $p < n$.

Forward-stepwise selection is basically the reverse of the backward-stepwsie method. More particularly, one starts with the model containing only an intercept, the variable that leads to the largest improvement is added, and this process is repeated until the performance

criterion can not be improved any further. The main advantage of this method over backward-stepwise selection is that it is generally less variable. On the other hand, this method usually leads to a bigger bias. Another limitation of forward-stepwise subset selection is that it fails if, in order to minimize some selection criterion, two predictors need to be included at the same time.

5.4.3 Univariate pre-screening

Underlying idea

Univariate pre-screening is somewhat related to forward-stepwise regression. Just as in forwar-stepwise selection we start by fitting the p models including an intercept and one predictor. In the second step a final model is fitted by using all predictors for which the corresponding model from the first step met a certain criteria, e.g. a significant p-value.

An advantage of this method is that, compared to the other subset selection methods that were discussed, it is computationally efficient. An important disadvantage of univariate pre-screening is that, because of its univariate nature, the correlation between predictors is ignored. Hence, highly correlated predictors will often be included in the final model. It is interesting to note that the multiple testing problem is particularly clear when using this method. However, since we know the exact number of tests it is quite easy to control some multiple testing error rate instead of the type 1 error. Since we will not use standard univariate pre-screening we will not discuss methods to control the multiple error rate. For two methods to control a multiple testing error rate we refer the interested reader to e.g. Holm 1979 or Benjamini and Hochberg 1995.

Taking the correlation into account

In ecological research a variation on univariate pre-screening that tries to take into account the correlation is regularly used, e.g. Cord et al. 2014. The method, which we will call select07, selects one variable from each set of highly correlated variables. The algorithm works as follows:

1. Make a set A of all variables.
2. Calculate all correlations.
3. For each pair of variables with $|r| > 0.7$ we fit a univariate model.
4. For each pair selected in step 3. we remove the worst[1] performing variable from A.
5. Fit a final model that includes all variables left in the set A.

An implementation of this method was provided by Dormann et al. 2013. In this implementation both GAMs and logistic regression can be used and the performance of the univariate models is measured by the AIC value. Finally we note that using $|r| > 0.7$ is quite arbitrary, however this threshold seems quite popular in SDM.

[1] As usual, different performance measures can be used, e.g. AIC, p-values, etc.

5.5 Taking the scale hierarchy into account

Thuiller, Araújo, and Lavorel 2004 used a hierarchical model selection approach. More specifically, they investigated the effect of adding land-cover data to models build with climate data. To do this they used the following three steps:

1. A model with only climate variables is constructed by using a stepwise selection method.

2. Regress the residuals of the climate model on the land-cover variables and use a stepwise selection method to select the most influential variables.

3. Build a new model with the selected climate and land-cover variables.

Although the focus in the article was on stepwise regression techniques, the same procedure can be applied in combination with other variable selection methods. However, this hierarchical approach has a clear disadvantage, it cannot consider interactions between climate and land-cover variables. Furthermore, the obtained solution will most often be sub-optimal compared to using a selection procedure with all the variables. Hence we see little reason to consider this approach any further.

5.6 Meaningful combinations of classification and

For each method of Chapter 4 a "vanilla" model will be considered. The vanilla logistic regression and ANN model use all predictors except the agriculture land cover class and BIO7. These two predictors are removed to avoid computational problems associated with design matrices that are not of full rank. For the vanilla GAM all predictors are used together with GCV to select the smoothness parameters. The vanilla MaxEnt model is the standard MaxEnt model with the default parameters (Phillips and Dudík 2008). Finally, the vanilla version of boosted regression trees is just the standard implementation.

Next to the vanilla models ten combinations of classification algorithms and methods to reduce the number of variables will be considered, see Table 5.1. Combinations that are not included were often not computationally feasible to implement, e.g. combing kernel PCA and GAMs, or not meaningful, e.g. L_1 regularization combined with boosted regression trees. Furthermore, following the remarks in Section 5.2.3 when PCA or kernel PCA are used in combination with presence-only data three different version will be considered:

1. A version where the PCA is performed on the predictor values associated with background and presence locations.

2. A version where the PCA is performed only on the predictors values associated with background locations.

3. A version where the PCA is performed only on the predictors values associated with presence locations.

In total 20 methods are used when presence-only data is considered and 14 when presence-absence data is considered.

[2] Regularization of the "wiggliness" instead of the coefficients, see Section 4.2.2

[3] Using five fold CV to select the L_1 regularization parameter instead of using the default parameter.

	PCA	kernel PCA	L_2 penalty	L_1 penalty	stepwise selection	select07
Logistic regression	✓	✓	✓	✓	✓	✓
Additive logistic regression			✓[2]			✓
Artificial neural networks			✓			
Boosted regression trees						
MaxEnt				✓ [3]		

Table 5.1: Table with the combinations of classification and dimensionality reduction techniques that are considered.

Chapter 6

Applications

6.1 Introduction

The data-sets were split into a training and a test set. The training data consists of $^3/_4$'th of the data while the test set contains the other $^1/_4$'th.

6.2 Implementations and tuning parameters of the methods

For the majority of the methods described in Chapters 5 and 4 some decisions regarding their model specification and tuning parameters have to be made.
In order to have a nice baseline method the logistic model was fit by using only linear terms. The logistic regression methods combined with the lasso or ridge were fit by using a cubic polynomial expansion of the predictors. To limit the amount of predictors and computational cost no interaction terms were considered.

It was decided to implement our own version of PCA logistic and KPCA logistic regression. This was mainly done to investigate the remarks made in Section 5.2.3. Both implementations allow us to specify on which indices the (K)PCA should be based on. Furthermore 5-fold-CV is used to select the an optimal number of principal components where always the components corresponding to the x highest eigenvalues. To allow some non-linearity in the PCA logistic regressin implementation we also allow for polynomial expansions in the PCs of up to the third degree (ignoring interaction effects). CV is used to select the optimal polynomial expansion. This expanding of the PCs into a polynomial is not done when KPCA is used, the reason is that KPCA already implicitly expands the original data into a new non-linear expansion. To limit the computational costs it was decided to limit the number of components to at most 75 in the KPCA implementation. Furthermore, when using KPCA together with logistic regression the KPCA will be performed on at most 500 observations, if necessary these are randomly selected from the relevant observations.

In the MaxEnt implementation where CV was used to select the regularization parameters considered were $\chi \times \lambda_{def}$ with λ_{def} the default MaxEnt parameter and $\chi \in \{0.1, 0.5, 1, 2, 10\}$.

The tuning variables and of the ANN method are the number of neurons and the in

the hidden layer $\in \{5, 10, 20, 40, 60\}$. Following the suggestions of Venables, Ripley, and Venables 2002 we optimize the weight decay parameter over $\{0.1, 0.01, 0.001, 0.0001\}$. Furthermore instead of simply averaging neural networks it was decided to use bagging which has, next to the advantages mentioned in Section **??**, the advantage that it tends to prevent over-fitting. The only tuning parameter that is used in the vanilla ANN model is the number of neurons in the hidden layer.

The tuning parameters of our gradient boosted decision trees are the number of trees $\in \{100, 500, 1000, 2000\}$, the interaction depth $\in \{1, 3, 5, 7\}$, and the amount of shrinkage $\in \{0.1, 0.01, 0.001, 0.0001)\}$. The choice of these values was mainly inspired by Elith, Leathwick, and Hastie 2008.

6.3 AUC as a measure of classification performance in SDM

To measure the performance of the classification method the area under the receiver operating curve (AUC) values will be used. Using the AUC to measure the performance of SDMs has been criticized (Jiménez-Valverde 2012; Lobo, Jiménez-Valverde, and Real 2008). However, most of the issues that are usually raised are not of particular interest in our case. Before studying the AUC values it is interesting to note that there are three main sources of randomness in the calculation of the mean AUC values across the different species.

First of all, in our set-up we could consider the species that are considered as drawn from a pool of all species in the databases. Of course this assumption is violated since we selected the species conditional on certain characteristics.

Secondly, the parameters in the classifications methods are estimated by using the training set. The elements of the training set can be seen sample of presence, and background or absence points from the corresponding distributions. The predicted variables, and the AUC, are therefore also dependent on the sample of points used.

Thirdly, the AUC is calculated on the test set which again consists of a sample of random presence, and background or absence points from the corresponding distributions. Thus even if the classifier was fixed we would get different AUC values if we use different test sets.

It is well known that the AUC statistic is equivalent to the Mann-Whitney U test statistic (Hanley and McNeil 1982) and hence the theory surrounding the Mann-Whitney U test statistic can provide us with standard errors (SE), asymptotic distributions, etc. However, these results only hold if the classifier is assumed to be fixed and hence ignores the randomness of the methods themselves and the species sampling scheme. A rigorous way to inspect the distribution of the AUC values would be to use a bootstrap method to construct e.g. confidence intervals. In our situation this is however computationally infeasible. In the end we opted to only report standard deviations (SD) and the mean values.

6.4 Presence-only data

6.4.1 Results

All the methods described in Sections 5.6 and 6.2 were fitted upon the GBIF data. Summary measures of the AUC values can be found in Table 6.1 and Figure 6.1.

	All variables		Bioclimatic variables		Difference	
Method	Mean AUC	SD	Mean AUC	SD	Mean AUC	SD
Logistic: vanilla	0.934	0.045	0.916	0.070	0.019	0.053
Logistic: backward	0.772	0.266	0.934	0.035	-0.162	0.262
Logistic: forward	0.661	0.242	0.599	0.277	0.062	0.330
Logistic: PCA	0.891	0.077	0.881	0.087	0.010	0.070
Logistic: presence PCA	0.838	0.172	0.887	0.090	-0.049	0.101
Logistic: background PCA	0.893	0.079	0.890	0.077	0.003	0.054
Logistic: kernel PCA	0.933	0.050	0.922	0.058	0.011	0.030
Logistic: presence kernel PCA	0.947	0.033	0.927	0.052	0.020	0.040
Logistic: background kernel PCA	0.937	0.048	0.935	0.039	0.002	0.028
Logistic: lasso	0.938	0.041	0.924	0.057	0.014	0.037
Logistic: ridge	0.930	0.047	0.908	0.063	0.022	0.039
Logistic: select07	0.938	0.040	0.915	0.058	0.023	0.032
GAM: auto selection	0.928	0.072	0.935	0.033	-0.008	0.069
GAM: vanilla	0.912	0.086	0.936	0.038	-0.024	0.090
GAM: select07	0.940	0.050	0.938	0.036	0.002	0.039
MaxEnt	0.921	0.055	0.932	0.041	-0.011	0.028
MaxEnt Vanilla	0.899	0.099	0.919	0.049	-0.020	0.076
ANN	0.946	0.046	0.948	0.032	-0.003	0.040
ANN: vanilla	0.926	0.058	0.934	0.039	-0.008	0.048
GBM	0.958	0.032	0.944	0.033	0.014	0.030

Table 6.1: Summary of the AUC values of the different classifiers fitted on the presence-only data.

First of all, few rigorous tests will be performed in this section. This because given the small sample size and the fact that the AUC values are quite close the power of such tests, using the usual 5% significance level, is very low. Because of this the discussion is rather short and mainly used as a preliminary study which generated interesting research questions for the simulation study of Chapter 7.

It is interesting to note that logistic regression performs quite well. Furthermore, the SDs of the stepwise selection methods are relatively large. This might indicate that these methods are inherently variable and hence lead to unstable predictions. It is interesting to note that while the performance of forward stepwise selection is bad when all variables are considered this is not the case when only the bioclimatic variables are considered. This could be explained by the fact that the backward selection method becomes a lot more variable when variables with little explanatory power are added. However, since this is a variable selection technique this is unwanted behaviour. The MaxEnt implementation that uses the default settings performs worse than the implementation in which the penalization

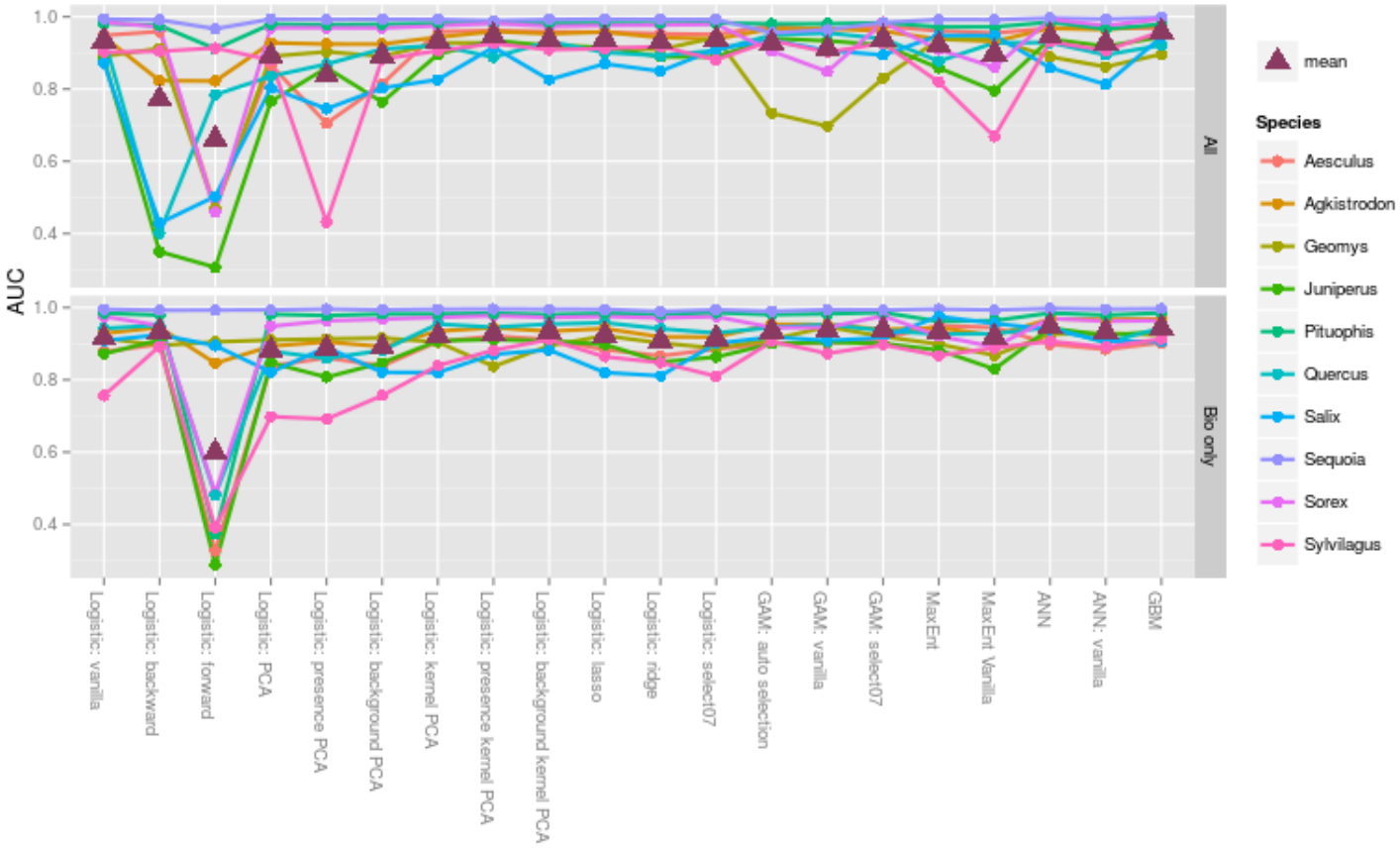

Figure 6.1: AUC values of the different classifiers fitted on the presence-only data.

parameter is selected by CV. Perhaps the most interesting is that GBM and ANNs seem to consistently lead to high AUC values with low standard deviations. Given our results it seems like MaxEnt might be overused throughout the standard deviations and at least some other machine learning methods should at least be considered before opting for MaxEnt.

In light of the discussion in Section 5.2.3 two Wilcoxon signed-rank tests were performed to test whether there is a significant difference between performing the (K)PCA on the background points or on the presence points. The p-value corresponding with the test for the PCA (resp. KPCA) is 0.32 (resp. 0.11) and hence accept the null-hypothesis of no difference between the methods.

To test whether or not using only the bioclimatic variables has a profound effect a Wilcoxon signed-rank test was performed for each method. These tests are of interest because on one hand they provide information on the usefulness of non-bioclimatic variables, i.e. is there an increase in classification performance. Also the a decrease in classification performance is of interest to detect, this could be interpreted as adding irrelevant predictors which lead to over-fitting. Even before performing a multiple testing correction none of the tests were significant at the 5% level. Hence we conclude that in our limited study there's no difference between using only the bioclimatic variables or all variables.

6.5 Presence-absence data

Although a variation of the MaxEnt method can be used for presence-only data this is nearly never done. Furthermore, instead of fitting three different variations of (K)PCA logistic regression it was decided to only use the standard approach. This was done to speed up the computations, and unlike what was he case in the presence-only scenario the absence points are "real" observations. Since only five species were considered there are no statistically significant results in this section. The inspection is mainly done to make sure that the different methods show approximately have the same performance characteristics for both presence-only and presence-absence data. The results can be found in Section **??**. Finally, because there are at between 42029 and 287860 plot locations available the fitting of the models leads to computational problems. A solution to these problems is proposed in Section 6.5.1

6.5.1 Case-control sampling

Because of computational considerations it was decided to use a subset of the presence-absence data to fit the models upon. More particularly, the subsample consists of all the presence points and 5000 randomly sampled absence points. This is clearly a form of case-control sampling and the effect of doing this has been studied to some extent (King and Zeng 2001).

In the case of logistic regression some algebra along the lines of Equation 4.3 shows that only the intercept is affected by the sub-sampling (King and Zeng 2001). Although the underlying coefficients stay the same the variance of the estimators should increase slightly. Because we only subsample the absence points this increase should be somewhat limited. Intuitively it makes sense that a rare presence point contains more information about the coefficients than one of the abundant absence points. In the case of logistic regression this can rigorously be deduced by observing that a term of the form $P(Y = 1|\boldsymbol{X} = \boldsymbol{x})\left[1 - P(Y = 1|\boldsymbol{X} = \boldsymbol{x})\right]$ turns up in the variance expression. Usually this term is quite a lot smaller for the abundant cases and hence leaving some out does not lead to a huge increase of the variance.

Although the previous paragraph focusses on logistic regression the same reasoning can be applied to the other methods considered. Finally, in the last two decades there has been quite some research on imbalanced data-sets and perhaps more efficient ways of dealing with the imbalance exist, see e.g. Chawla 2005 for an overview of other methods.

6.5.2 results

The mean and and the standard deviations can be found in Table 6.2 and Figure 6.2 shows a plot of the estimated AUC values.

It is readily seen that the logistic regression model performs quite well, i.e. it leads to relatively high average AUC values and its SD is quite low. The stepwise methods both have large SD and have the lowest average AUC values. The GBM and ANN methods are the two methods with the largest average AUC and both have relatively small SD. Unlike in the presence-only scenario the logistic07 methods has quite a large variance and

	All variables		Bioclimatic variables		Difference	
Method	Mean AUC	SD	Mean AUC	SD	Mean AUC	SD
Logistic: vanilla	0.961	0.027	0.960	0.028	0.001	0.001
Logistic: backward	0.682	0.301	0.656	0.318	0.026	0.345
Logistic: forward	0.774	0.284	0.741	0.256	0.034	0.046
Logistic: PCA	0.913	0.084	0.913	0.084	0.000	0.000
Logistic: kernel PCA	0.950	0.052	0.950	0.052	0.000	0.000
Logistic: lasso	0.960	0.042	0.960	0.042	0.000	0.000
Logistic: ridge	0.957	0.044	0.957	0.044	0.000	0.000
Logistic: select07	0.756	0.239	0.756	0.239	0.000	0.000
GAM: auto selection	0.925	0.098	0.925	0.098	0.000	0.000
GAM: vanilla	0.877	0.117	0.877	0.117	0.000	0.000
GAM: select07	0.948	0.062	0.948	0.062	0.000	0.000
ANN	0.973	0.020	0.972	0.021	0.001	0.002
ANN: vanilla	0.962	0.036	0.969	0.029	-0.007	0.014
GBM	0.977	0.014	0.977	0.014	0.000	0.000

Table 6.2: Summary of the AUC values of the different classifiers fitted on the presence-absence data.

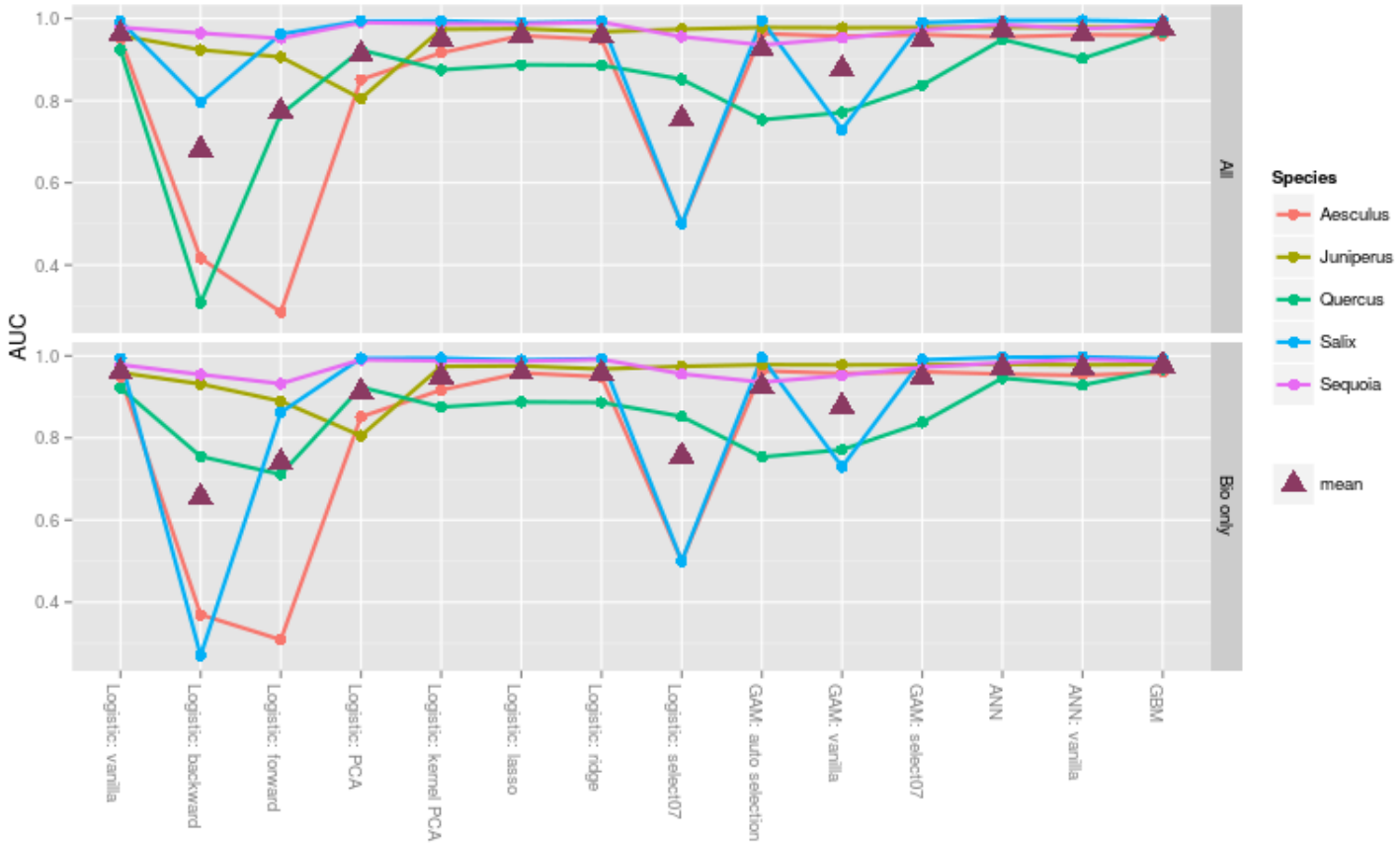

Figure 6.2: AUC values of the different classifiers fitted on the presence-absence data.

performs quite badly. Finally, the differences between using the bioclimatic or all variables are extremely sall

6.6 Discussion

Chapter 7

Simulation study

7.1 Introduction

In order to study the methods in more depth a simulation study was performed. The main advantage of this simulation study over using real species, as was done in Chapter 6, is that we can decide on which variables make up the model. Because of the usual non-linear shape of the response-functions, the sampling design, ... it is quite hard to set up a simulation study. Luckily the VIRTUALSPECIES package (Leroy et al. 2014, 2015) provides a simulation framework that includes important characteristics of species' distributions.

7.2 Overview of the virtualspecies package and the simulation set-up

First of all, the VIRTUALSPECIES package generates a suitability raster, i.e. a raster containing high values for suitable habitat and vice versa, based upon a set of provided rasters. Since the package imposes no restrictions on the function that generates the suitability the suitability raster has to be converted into a probability of occurrence map. This can be done by using e.g. the logit function to map $\mathbb{R} \to [0, 1]$. Once a probability map is obtained a raster containing 1 where the simulated species is present and 0 otherwise can be obtained by sampling cells with a probability proportional to the probability of the occurrence map. Finally from this raster we can obtain a presence-only sample by using a drawing random points within the cells that have a value of 1 in the species distribution raster.

To facilitate the computations it was decided to select one area in which several random species were generated. In order to make sure that the environmental conditions can fluctuate within the distribution of one generated species to the next it was required that the selected area has to contain different types of habitat types. The obvious choice would be to use the contiguous 47, however this has a huge computational demands. In the end it was decided to to perform the simulation study on data contained within a rectangle that approximately corresponds to Washington state, see Figure 7.1. Washington state has a big precipitation and temperature gradient because of the rain shadow created by the Cascade Range. This is also reflected in the fact that the types of habitat within this extent range from temperate rainforests, e.g. the Hoh Rainforest, to steppe- and

dessert-like areas in Southeastern Washington.

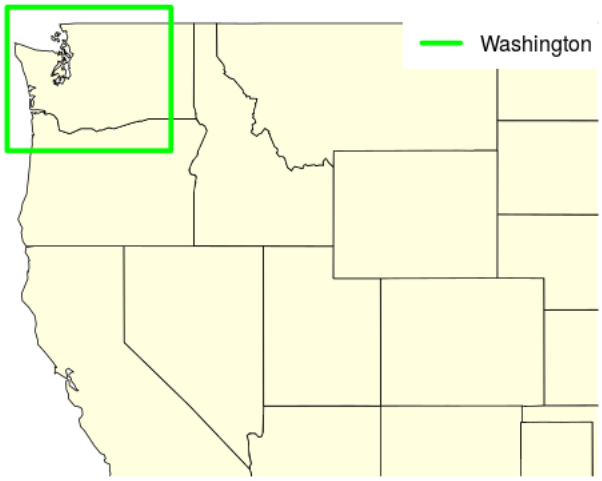

Figure 7.1: Extent considered in the simulation study.

In order to study the generalizability of the methods it was decided that 5 different species would be simulated. Furthermore, to be able to investigate the sample to sample variability of each method 5 different presence-only samples were drawn for each simulated species. Since the mean (resp. median) number of observations of the presence-only datasets is 320.3 (resp. 133.5) it was decided to simulate 200 occurrence points for each simulated dataset.

In Section 6.3 the various sources of the variability in the AUC values were discussed. Given that the interest is mainly in the training sample to training sample variability it was decided to try to restrict the test sample to test sample variability to some extent. This can be done by taking a large test set. Using a sample of 1000 occurrence and 1000 background locations the standard error of the AUC value, conditional on the classifier, is ≈ 0.013. This should

7.3 Results

7.4 Discussion

Chapter 8

Conclusion

8.1

8.2 Future research

Although the results in this thesis give some indication of the performance of variable selection methods the number of studied species is too small to make conclusive recommendations. Investigating the performance of the different models for additional species, preferably also in different continents, should solidify the conclusions.

It seems that introducing new statistical techniques to SDM could lead to significant improvements over current practices. Given that it is quite hard to manually check whether observations are outliers or not, see Chapter 3. In the statistical literature methods that are not (heavily) influenced by outliers are called robust. Robust variations of GLMs and IPP models are already available (Assunção and Guttorp 1999; Cantoni and Ronchetti 2001). However, if robust version of GLMs are used in combination with background data it could be interesting to adapt the functions so that background points are never considered to be outliers. Furthermore, the connections between MaxEnt, IPP models, and GLMs should make it doable to extent the robust variations of GLMs and IPP models to construct a robustified MaxEnt method. Finally, also model selection is influenced by outliers recently some methods have become available that deal with this scenario (e.g. Müller and Welsh 2009)

In Section 6.5.1 subsampling presence-absence data was introduced. Instead of taking a completely random subsample of the absence points more advanced methods exist (King and Zeng 2001). Although the gain of these methods might turn out to be minimal it could open the door to using even more advanced methods that are as of yet too computationally intensive.

Even though we tried to study a number of different variable selection techniques a lot of different approach are available or being developed. Examples include the focused information criterion (FIC, Claeskens et al. 2003), elastic net (Zou and Hastie 2005), ... The FIC is a information criterion that can be adjusted for the goal of the study. E.g. a FIC can be constructed that focusses on variable selection with as goal the predictions of occurrence probability in a few select locations. The elastic net is a new penalization technique that combines the L_2 and L_1 penalties and has been shown to perform quite well.

Bibliography

Agresti, Alan (2013). *Categorical data analysis*. 3rd ed. Wiley series in probability and statistics 792. Hoboken, NJ: Wiley.

Anderson, James R. et al. (1976). *A land use and land cover classification system for use with remote sensor data*. USGS Numbered Series 964.

Assunção, Renato and Peter Guttorp (1999). "Robustness for Inhomogeneous Poisson Point Processes." en. In: *Annals of the Institute of Statistical Mathematics* 51.4, pp. 657–678.

Barbet-Massin, Morgane et al. (2012). "Selecting pseudo-absences for species distribution models: how, where and how many?" en. In: *Methods in Ecology and Evolution* 3.2, pp. 327–338.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Bishop, Christopher M. (1995). *Neural networks for pattern recognition*. Oxford : New York: Clarendon Press ; Oxford University Press.

Cantoni, Eva and Elvezio Ronchetti (2001). "Robust Inference for Generalized Linear Models." In: *Journal of the American Statistical Association* 96.455, pp. 1022–1030.

Chawla, Nitesh V (2005). "Data mining for imbalanced datasets: An overview." In: *Data mining and knowledge discovery handbook*. Springer, pp. 853–867.

CIAT-CSI SRTM. `http://srtm.csi.cgiar.org/`.

Claeskens, Gerda et al. (2003). "The Focused Information Criterion." In: *Journal of the American Statistical Association* 98.464, pp. 900–945.

Colwell, Robert K. and Thiago F. Rangel (2009). "Hutchinson's duality: The once and future niche." en. In: *Proceedings of the National Academy of Sciences* 106.Supplement 2, pp. 19651–19658.

Cord, Anna F. et al. (2014). "Remote sensing data can improve predictions of species richness by stacked species distribution models: a case study for Mexican pines." en. In: *Journal of Biogeography* 41.4, pp. 736–748.

Daly, Christopher et al. (2002). "A knowledge-based approach to the statistical mapping of climate." In: *Climate Research* 22.2, pp. 99–113.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

DiMiceli, C.M. et al. *Annual Global Automated MODIS Vegetation Continuous Fields (MOD44B) at 250 m Spatial Resolution for Data Years Beginning Day 65, 2000 - 2010, Collection 5 Percent Tree Cover*. University of Maryland, College Park, MD, USA.

Dormann, Carsten F. et al. (2013). "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance." en. In: *Ecography* 36.1, pp. 27–46.

Elith, J., J. R. Leathwick, and T. Hastie (2008). "A working guide to boosted regression trees." en. In: *Journal of Animal Ecology* 77.4, pp. 802–813.

Elith, Jane and John R. Leathwick (2009). "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." In: *Annual Review of Ecology, Evolution, and Systematics* 40.1, pp. 677–697.

Elith*, Jane et al. (2006). "Novel methods improve prediction of species' distributions from occurrence data." en. In: *Ecography* 29.2, pp. 129–151.

Fithian, William and Trevor Hastie (2013). "Finite-sample equivalence in statistical models for presence-only data." In: *The Annals of Applied Statistics* 7.4. arXiv: 1207.6950, pp. 1917–1939.

Franklin, Janet and Jennifer A. Miller (2009). *Mapping species distributions: spatial inference and prediction*. Ecology, biodiversity and conservation. Cambridge ; New York: Cambridge University Press.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2000). "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)." EN. In: *The Annals of Statistics* 28.2, pp. 337–407.

Friedman, Jerome et al. (2015). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 2.0-2.

Fry, Joyce A et al. (2011). "Completion of the 2006 national land cover database for the conterminous United States." In: *Photogrammetric engineering and remote sensing* 77.9, pp. 858–864.

Guisan, Antoine, Thomas C Edwards Jr, and Trevor Hastie (2002). "Generalized linear and generalized additive models in studies of species distributions: setting the scene." In: *Ecological Modelling* 157.2–3, pp. 89–100.

Guisan, Antoine and Wilfried Thuiller (2005). "Predicting species distribution: offering more than simple habitat models." en. In: *Ecology Letters* 8.9, pp. 993–1009.

Guisan, Antoine and Niklaus E. Zimmermann (2000). "Predictive habitat distribution models in ecology." In: *Ecological Modelling* 135.2–3, pp. 147–186.

Hanley, J A and B J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Hastie, Trevor and Werner Stuetzle (1989). "Principal Curves." In: *Journal of the American Statistical Association* 84.406, pp. 502–516.

Hastie, Trevor and Robert Tibshirani (1990). *Generalized additive models*. 1st ed. Monographs on statistics and applied probability 43. London ; New York: Chapman and Hall.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer series in statistics. New York, NY: Springer.

Heikkinen, Risto K. et al. (2007). "Biotic interactions improve prediction of boreal bird distributions at macro-scales." en. In: *Global Ecology and Biogeography* 16.6, pp. 754–763.

Hijmans, Robert J. (2015). *raster: Geographic Data Analysis and Modeling*. R package version 2.4-20.

Hijmans, Robert J. et al. (2015). *dismo: Species Distribution Modeling*. R package version 1.0-12.

Hof, Anouschka R., Roland Jansson, and Christer Nilsson (2012). "The usefulness of elevation as a predictor variable in species distribution modelling." In: *Ecological Modelling* 246, pp. 86–90.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure." In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70.

Homer, Collin et al. (2007). "Completion of the 2001 national land cover database for the counterminous United States." In: *Photogrammetric Engineering and Remote Sensing* 73.4, p. 337.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators." In: *Neural Networks* 2.5, pp. 359–366.

Jiménez-Valverde, Alberto (2012). "Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling." en. In: *Global Ecology and Biogeography* 21.4, pp. 498–507.

Karatzoglou, Alexandros et al. (2004). "kernlab - An S4 Package for Kernel Methods in R." In: *Journal of Statistical Software* 11.1, pp. 1–20.

King, Gary and Langche Zeng (2001). "Logistic Regression in Rare Events Data." en. In: *Political Analysis* 9.2, pp. 137–163.

Kuhn, Max et al. (2015). *caret: Classification and Regression Training*. R package version 6.0-58.

Leroy, Boris et al. (2014). *virtualspecies: Generation of Virtual Species Distributions*. R package version 1.0.

— (2015). "virtualspecies, an R package to generate virtual species distributions." en. In: *Ecography*, n/a–n/a.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real (2008). "AUC: a misleading measure of the performance of predictive distribution models." en. In: *Global Ecology and Biogeography* 17.2, pp. 145–151.

Marra, Giampiero and Simon N. Wood (2011). "Practical variable selection for generalized additive models." In: *Computational Statistics & Data Analysis* 55.7, pp. 2372–2387.

Marx, Brian D. (1996). "Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression." In: *Technometrics* 38.4, pp. 374–381.

McCullagh, Peter and John Ashworth Nelder (1999). *Generalized linear models*. eng. 2. ed., [Nachdr.] Monographs on statistics and applied probability 37. London: Chapman & Hall.

Merow, Cory et al. (2014). "What do we gain from simplicity versus complexity in species distribution models?" en. In: *Ecography* 37.12, pp. 1267–1281.

Møller, Jesper and Rasmus P. Waagepetersen (2007). "Modern Statistics for Spatial Point Processes*." en. In: *Scandinavian Journal of Statistics* 34.4, pp. 643–684.

Müller, Samuel and A. H. Welsh (2009). "ROBUST MODEL SELECTION IN GENERALIZED LINEAR MODELS." In: *Statistica Sinica* 19.3, pp. 1155–1170.

O'Connel, Barbara M et al. (2015). *The Forest Inventory and Analysis Database: Database Description and User Guide Version 6.0.2 for Phase 2*. U.S. Forest Service.

Oke, Oluwatobi A. and Ken A. Thompson (2015). "Distribution models for mountain plant species: The value of elevation." In: *Ecological Modelling* 301, pp. 72–77.

Parisien, Marc-André and Max A. Moritz (2009). "Environmental controls on the distribution of wildfire at multiple spatial scales." In: *Ecological Monographs* 79.1, pp. 127–154.

Pearce, Jennie L. and Mark S. Boyce (2006). "Modelling distribution and abundance with presence-only data." en. In: *Journal of Applied Ecology* 43.3, pp. 405–412.

Pearson, Richard G. and Terence P. Dawson (2003). "Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful?" en. In: *Global Ecology and Biogeography* 12.5, pp. 361–371.

Pearson, Richard G., Terence P. Dawson, and Canran Liu (2004). "Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data." en. In: *Ecography* 27.3, pp. 285–298.

Pebesma, Edzer et al. (2015). *sp: Classes and Methods for Spatial Data*. R package version 1.2-1.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire (2006). "Maximum entropy modeling of species geographic distributions." In: *Ecological Modelling* 190.3–4, pp. 231–259.

Phillips, Steven J. and Miroslav Dudík (2008). "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation." en. In: *Ecography* 31.2, pp. 161–175.

Pinzon, J, Molly E Brown, and Compton J Tucker (2005). "Satellite time series correction of orbital drift artifacts using empirical mode decomposition." In: *Hilbert-Huang transform: introduction and applications* 16.

PRISM Climate Group, Oregon State University created 4 Feb 2004. `http://prism.oregonstate.edu`. created 17 Nov 2015.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Renner, Ian W. and David I. Warton (2013). "Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology." en. In: *Biometrics* 69.1, pp. 274–281.

Ridgeway, Greg (2015). *gbm: Generalized Boosted Regression Models*. R package version 2.1.1.

Ripley, Brian and William Venables (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-12.

Ripley, Brian D. (2009). *Pattern recognition and neural networks*. eng. 1. paperback ed. 1997, reprinted 2009. Cambridge: Cambridge Univ. Press.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors." en. In: *Nature* 323.6088, pp. 533–536.

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller (1997). "Kernel principal component analysis." en. In: *Artificial Neural Networks — ICANN'97*. Ed. by Wulfram Gerstner et al. Lecture Notes in Computer Science 1327. Springer Berlin Heidelberg, pp. 583–588.

Segurado, Pedro and Miguel B. Araújo (2004). "An evaluation of methods for modelling species distributions." en. In: *Journal of Biogeography* 31.10, pp. 1555–1568.

Soberón, Jorge (2007). "Grinnellian and Eltonian niches and geographic distributions of species." en. In: *Ecology Letters* 10.12, pp. 1115–1123.

Soberón, Jorge and Miguel Nakamura (2009). "Niches and distributional areas: Concepts, methods, and assumptions." en. In: *Proceedings of the National Academy of Sciences* 106.Supplement 2, pp. 19644–19650.

Strubbe, Diederik and Erik Matthysen (2008). "Predicting the potential distribution of invasive ring-necked parakeets Psittacula krameri in northern Belgium using an ecological niche modelling approach." en. In: *Biological Invasions* 11.3, pp. 497–513.

Thuiller, Wilfried, Miguel B Araújo, and Sandra Lavorel (2004). "Do we need land-cover data to model species distributions in Europe?" en. In: *Journal of Biogeography* 31.3, pp. 353–361.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Tucker, Compton J et al. (2005). "An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data." In: *International Journal of Remote Sensing* 26.20, pp. 4485–4498.

Venables, W. N., Brian D. Ripley, and W. N. Venables (2002). *Modern applied statistics with S*. 4th ed. Statistics and computing. New York: Springer.

Vogelmann, James E et al. (2001). "Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources." In: *Photogrammetric Engineering and Remote Sensing* 67.6.

Ward, Gill et al. (2009). "Presence-Only Data and the EM Algorithm." en. In: *Biometrics* 65.2, pp. 554–563.

Warton, David I. and Leah C. Shepherd (2010). "Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology." EN. In: *The Annals of Applied Statistics* 4.3, pp. 1383–1402.

Whittingham, Mark J. et al. (2006). "Why do we still use stepwise modelling in ecology and behaviour?" en. In: *Journal of Animal Ecology* 75.5, pp. 1182–1189.

Wiens, John A. et al. (2009). "Niches, models, and climate change: Assessing the assumptions and uncertainties." en. In: *Proceedings of the National Academy of Sciences* 106.Supplement 2, pp. 19729–19736.

Wildlife Conservation Society - WCS and AU - Center for International Earth Science Information Network. *Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Influence Index (HII) Dataset (Geographic)*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Accessed 9 Nov 2015.

Wood, Simon (2015). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-7.

Wood, Simon N. (2006). *Generalized additive models: an introduction with R*. Texts in statistical science. Boca Raton, FL: Chapman & Hall/CRC.

Wood, Simon N. and Nicole H. Augustin (2002). "GAMs with integrated model selection using penalized regression splines and applications to environmental modelling." In: *Ecological Modelling* 157.2–3, pp. 157–177.

Woodward, F. I., G. E. Fogg, and U. Heber (1990). "The Impact of Low Temperatures in Controlling the Geographical Distribution of Plants [and Discussion]." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 326.1237, pp. 585–593.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net." en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Leuven Statistics Research Centre (LStat)
Celestijnenlaan 200 B, bus 5307
3001 HEVERLEE, BELGIË
tel. +32 16 377 111
www.kuleuven.be