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

Preface

lala

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 (e.g. Pearson and Dawson 2003; Pearson, Dawson, and Liu 2004), the impact of invasive species, 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 SDM are discussed.

The data-sets and variables that will be used throughout this thesis are described in Chapter 3. The variables that are used in this thesis describe either the climate at a certain location or are derived from remotely sensed products. 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 4.1 we focus on using data-sets that consist of locations where the species was either present or absent. In this case standard classification methods can be

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

utilized. However, often the data-set only includes occurrence locations, for example data-sets from natural history musea or citizen science projects are usually of this type. To use occurrence only data the classical classification algorithms form Section 4.1 can be adapted, this is done in Section 4.2.1. Another approach is to use one of the algorithms specifically constructed for presence only data, one of these is introduced in Section 4.2.2.

In practice a large part of building SDMs consists of variable selection. The goal of this thesis is to investigate the performance of multiple model selection methods in settings that are representative of what could be expected in practice. In Chapter 5 we give an overview of often used methods to deal with large magnitudes of predictors. More particularly, we will introduce:

- Regularization.
- Step-wise selection.
- Dimensionality reduction of the explanatory variables.
- So-called "folklore" methods.

Finally, although we will introduce the most important concepts and some applications of species distribution modeling it is not the goal of this thesis to describe every aspect in detail. Instead we refer to Miller and Franklin 2002 who gave an overview of the field of species distribution modeling. Other introductory material include Guisan and Zimmermann 2000, Guisan and Thuiller 2005, and Elith and Leathwick 2009. 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. We can, non-rigorously, define the ecological niche of a species as the set of environmental conditions where its reproduction rate is larger 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 desription we refer the interested reader to Soberón 2007 and, 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 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.

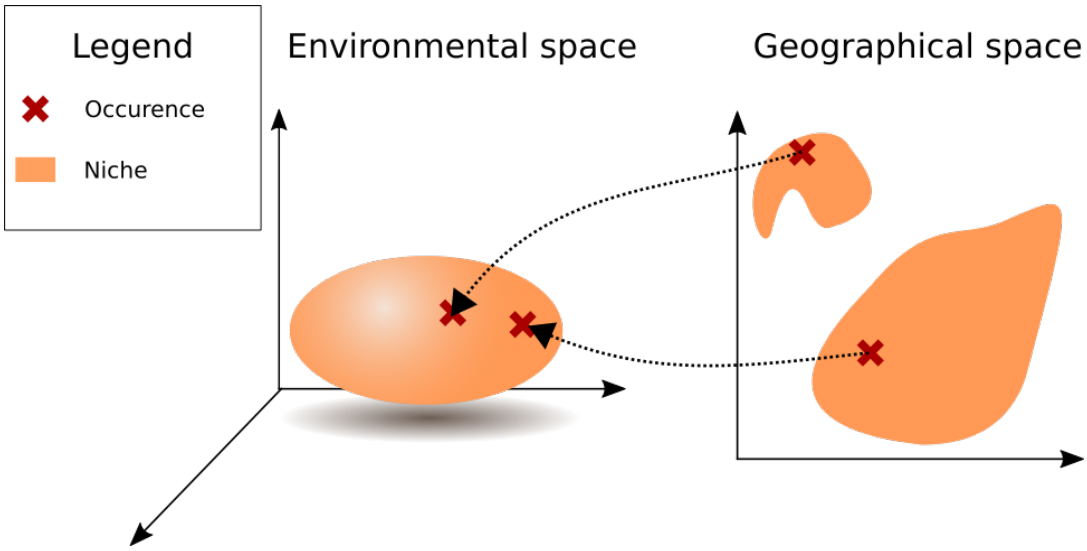

Figure 2.1: Visualization of the duality between environmental and geographical space.

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.

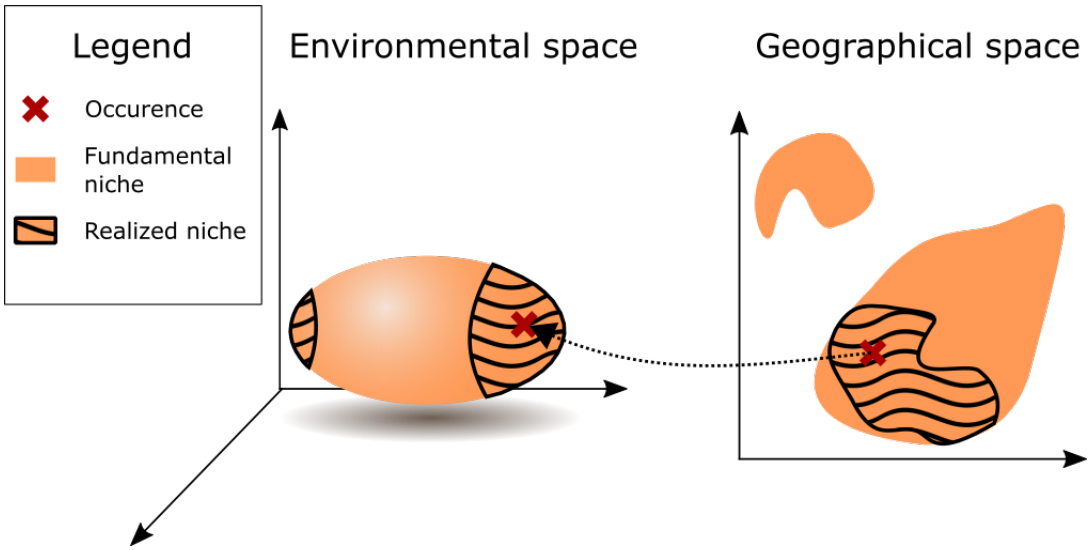

Figure 2.2: Visualization of the difference between the fundamental and realized niche.

2.3 Implicit assumptions when building and using species distribution modles

Before applying SDMs in practice we have to realize that there are some important underlying assumptions. We will only describe a few of these assumptions. 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.

By definition every observation in the field belongs to the realized niche. SDMs are therefore models of the realized niche. If we want to use a certain SDM to e.g. predict areas prone to invasive species, we implicitly assume 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 we use data to build a model of the realized niche we have to assume that the observed data-points are representative of the niche. In practical settings this is often not the case. For example, due to climate change tree species might be found in regions where the current environmental conditions are not included in its niche. The niche of a species usually evolves over time. This is another reason of why the observations may not be representative of the niche we are modelling. Hence, we have to assume that historical records are representative of the niche that is being modelled.

Another assumption 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 (Heikkinen et al. 2007).

Chapter 3

Data commonly used in species distribution models

In total 34 predictors are used. These predictors were selected because they are widely available and are often used in SDMs. Using 34 predictors to build a species distribution model is rather

3.1 Predictor data

To process the spatial data `R` 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. Although this chapter is only a small piece of the thesis the data-preparation is the most labour intensive part of it.

3.1.1 Vegetation Continuous Fields

The Vegetation Continuous Fields (VCF, DiMiceli et al.) data-set contains values between 0 and 100. These values lie in the interval $[0, 100]$ and correspond to the proportional tree cover of the cell. Some cells also contain values larger than 100 and these correspond to water or rasters with no available data. Since R has a NA value all the cells with values > 100 are set to NA. The raster is provided in the geogracphical coordinate system combined with the World Geodetic System 1984 datum (GCS_WGS84). The resolution of the VCF raster is 0.002083 decimal degrees.

3.1.2 Bioclimatic variables

The bioclimatic variables are a set of variables which 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.

Variable name	Explanation
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 ()
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Table 3.1: Explanation of the bioclimatic variables.

It is clear that the BIO05, BIO6, and BIO7 variables are linear 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. This will be discussed in more detail in Section 5.5.

The monthly temperature and precipitation data was obtained from the PRISM database (PRISM Climate Group). The PRISM rasters have a grid cell size of 0.00833 degrees and the rasters are provided in the GCS_WGS84 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 it uses the average of the minimum and maximum temperature whenever the calculations require the mean temperature. Our adaptation does use the mean temperature layers and should be slightly more realistic.

3.1.3 Normalized Difference Vegetation Index

The Normalized Difference Vegetation Index (NDVI) is a measurement of the amount of vegetation. It is based on measurements of the reflectance of the infra-red and the near infra-red region. The NDVI takes on values between -1 and 1 and high values correspond with live green vegetation. The NDVI raster used in this thesis originates form the Global Inventory Modeling and Mapping Studies (GIMMS) and is provided by the University of Maryland Global Land Cover Facility (Pinzon, Brown, and Tucker 2005; Tucker et al. 2005). This database contain 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 the semi-monthly rasters were combined such that monthly rasters were obtained. To obtain an average monthly NDVI raster for each month the 24 NDVI rasters were averaged. The 12 resulting NDVI rasters are then used to calculate a minimum, maximum, and mean NDVI raster.

3.1.4 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. The use of elevation in SDM is somewhat contested, see e.g. Hof, Jansson, and Nilsson 2012 and 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, however, to keep the amount of data in this thesis handleable we will restrict ourselves to the original raster.

3.1.5 Land cover

The land cover data is created from the National Land Cover Databases (NLCD) provided by the US Geological Survey. The NLCD are 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 class rasters with a cell size of 0.04166 decimal degrees of the years 2001, 2007, and 2011 were created. In order to obtain one raster these were then averaged. The eight final rasters contain values in the interval $[0, 1]$ and correspond to percentage of the respective land-cover within the cell. It is interesting that the sum of these rasters equals one, hence there is a linear dependence between the variables.

3.1.6 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 is reprojected to the GCS_WGS84 spatial reference system and has a cell size of 0.00833 decimal degrees.

3.1.7 Preprocessing of the predictor data

In order to speed up computations and facilitate general GIS operations the rasters are, if necessary, reprojected to the GCS_WGS84 spatial reference system. The extent and resolution of the rasters is set to be equal to those of the DEM layer. 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 is preprocessed the rasters amount to over 300 GB of data.

3.1.8 Exploratory analysis of the predictor data

Since one would expect that the relationship between the variables is different in different geographical regions we defined four subregions of the contiguous United States. The regions and their defining bounding rectangles can be found in Figure 3.2.

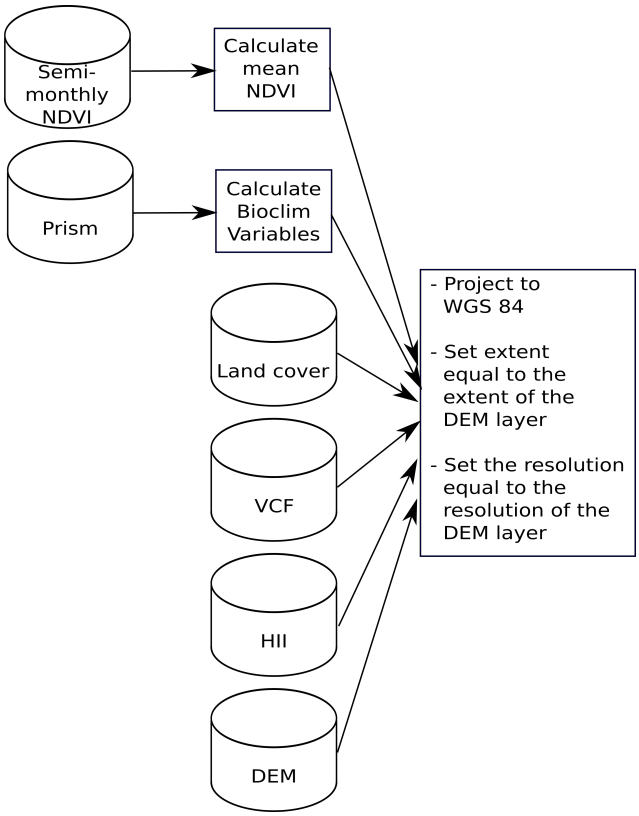

Figure 3.1: Visualization of the preprocessing of the raster data.

In order to check whether the relationship between the variables is different in different regions 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 seem to be rather stable. There are however also some 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 behaviour could be observed for the other regions.

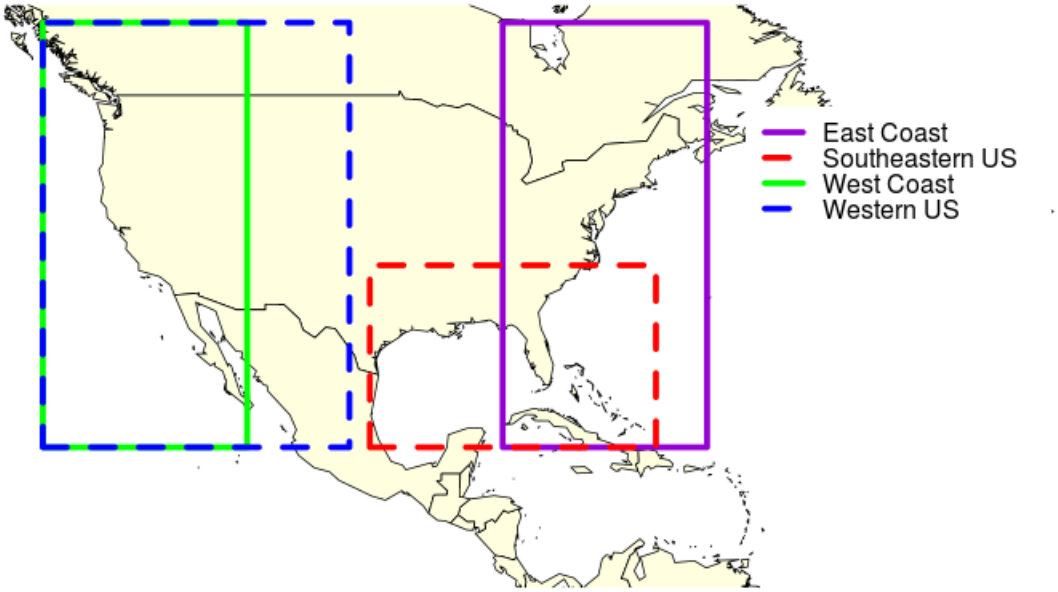

Figure 3.2: Regions of the contiguous US and their bounding rectangles.

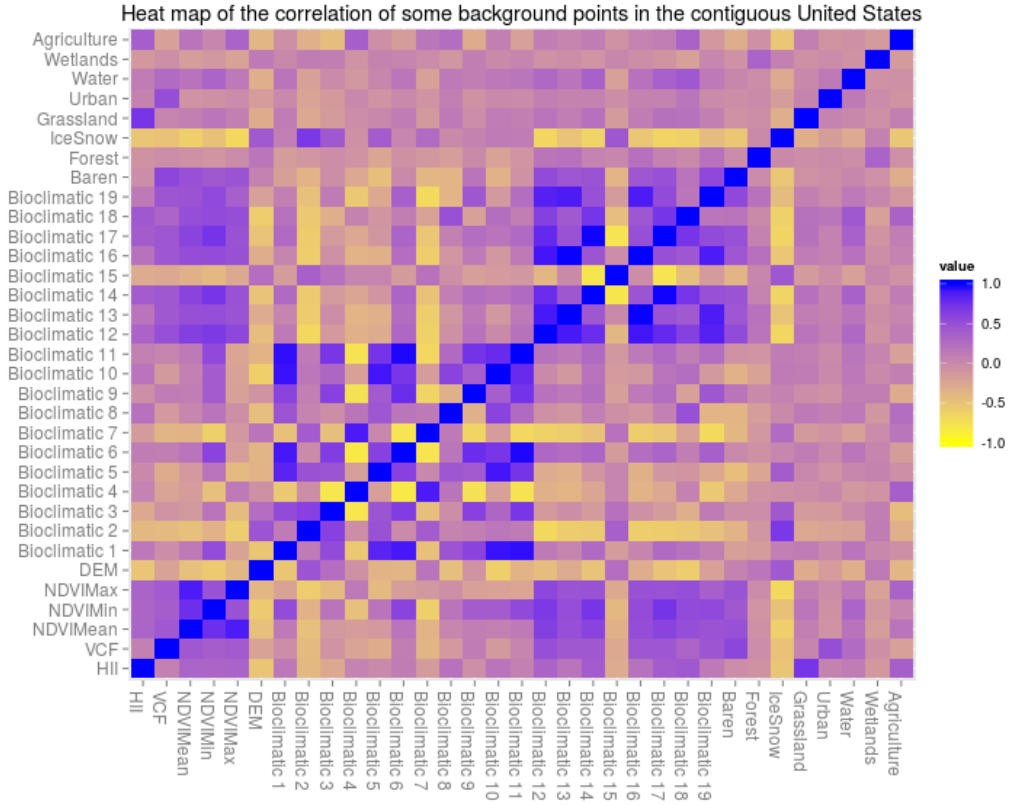

Figure 3.3: Heat map of the correlations in the US.

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

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 land cover variables should sum to a constant hence one would expect that the rank is smaller or equal to 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.

3.2 Outcome data

3.2.1 Species considered

The species that will be studied can be found in Table 3.2. These species were selected in such a way that different regions in the US are represented. This was done because, as we saw in Section 3.1.8, the relationship between the different predictors can be different in different regions. The extent of the species distribution is 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. Having species for which the distribution within the study area is smaller or larger should lead to a different relationship between the predictors and the distribution. Finally, we included five plant and five animal species. This 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.

3.2.2 Global Biodiversity Information Facility

The presence-only data originates from the Global Biodiversity Information Facility (GBIF) database. This database contains data from different other smaller occurrence only databases, e.g. 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. For example, 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 observations in botanical gardens etc. Furthermore GBIF data tends to contain a lot of duplicated observations. Hence, before using the data from GBIF some data-cleaning was performed. Finally, because the predictors were recorded quite recently we decided to restrict ourselves to observations obtained from the 1980's onward.

3.2.3 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. The locations that are reported 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 is sampled each year, hence a time-frame of at least 10 years means that each plot site should be sampled. The time-frame that we use is 2004-2014, Since the sampling in New Mexico, Oklahoma, and Wyoming started later than 2004 these states are undersampled. Furthermore, states in the Easter US tend to have a sample intensity larger than 10% and some plots were sampled multiple times. For the plots where this is the case we replace it with a new observation. If the plot contains the species of interest at least once the species is said to be present, otherwise it was absent in the plot. It might be interesting to use modelling methods that allow for a sampling design correct. However, this would lead is astray the sampling design will not be corrected for in this thesis.

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

3.2.4 data preparation

In order to build the necessary models the predictor values corresponding to the presence or absence locations need to be extracted. For some of the locations some rasters contain a NA value. These points are removed before the models are constructed. This might lead to some slight biases in the models but since the goals is not to construct perfect interpretable models but to inspect the predictive performance this should not lead to large problems.

3.3 Spatial scale

cite Fine-scale environmental variation in species distribution modelling ...

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 the study extent.

Chapter 4

Classification techniques

This chapter deals with the statistical foundations of species distribution modelling. Since it is impractical to list all the available methods that are used 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 we selected the methods we will shortly (re-)introduce.

4.1 Presence absence data

In this section some important and often used methods to classify binary data are review. First of all, the outcome, y, 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.

4.1.1 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

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

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

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

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:

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

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 and 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.

4.1.2 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 could be useful. When the distribution of the outcome belongs to the exponential family generalized additive models (GAMs) are one possible class of semi-parametric models. In the case of a Bernouilli 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 and 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 the minimizer of

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

in a class of sufficiently smooth functions 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 function converges to the linear logistic regression solution. Finally, natural cubic splines also have some computational advantages and are hence a good option to use as the smoothing functions in GAMs.

In practice we have to select the penalization parameters, this is usually done by using cross-validation. Because the fitting procedures for GAMs are often 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 is necessary to perform additional model selection to test whether or not to include a covariate. Although variable selection is the topic of Chapter 5 we mention that it is possible to introduce an extra penalization term which leads to an automatic variable selection technique (Marra and Wood 2011). This approach is implemented in the MGCV

R package (Wood 2015).

Finally, up till 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.

4.1.3 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.

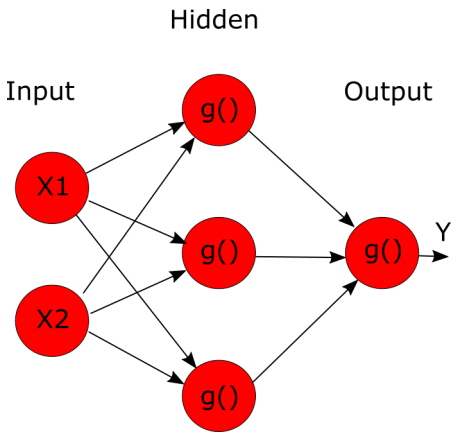

Figure 4.1: Visualization of a feed-forward neural network with one hidden layer.

From here on we will denote the vector of all the weights as $\boldsymbol{\gamma}$. The estimated weights $\widehat{\boldsymbol{\gamma}}$ are

obtained by minimizing a loss function $L(\boldsymbol{y}; \boldsymbol{\gamma})$:

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

If we denote the predicted values given some $\boldsymbol{\gamma}$ by $\hat{y}_i(\boldsymbol{\gamma})$ we can specify some loss functions. In classification problems often either the squared loss, $L(\boldsymbol{y}; \boldsymbol{\gamma}) = \sum_i (y_i - \hat{y}_i(\boldsymbol{\gamma}))^2$ or the cross-entropy, $L(\boldsymbol{y}; \boldsymbol{\gamma}) = -\sum_i [y_i \log\{\hat{y}_i(\boldsymbol{\gamma})\} + (1 - y_i) \log\{\hat{y}_i(\boldsymbol{\gamma})\}]$ is used. To minimize the loss criterion, often backpropagation (Rumelhart, Hinton, and Williams 1986) is used in combination with an numerical optimization algorithm e.g. steepest descent.

maybe fix the notation so that it's more in line with the previous chapters.

It is easy to see that logistic regression can also 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.

The main advantage of neural networks 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 5.2).
- The fitted parameters can be sensitive to the initial weights that are used at the first step of the numerical optimization algorithm.
- 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 5.2.1. Averaging networks boils down to fitting the same network structure but using different initial values and then combining these (Ripley 2009). The CARET package (Kuhn et al. 2015) provides an implimentation of averaged neural networks based upon the NNET package (Ripley and Venables 2016).

4.1.4 Tree based methods

This section introduces some tree based methods. First of all, decision trees are explained. 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 a 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. A visualization of a partitioning obtained bys using a classification tree is given in Figure 4.2.

Unless we specify a stopping criteria this approach would settings lead to over-fitting. More particularly we would end up with as many regions as there are observations. 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. Finally, we note that there are many variations to the algorithm sketched above.

discuss different loss functions that are often used, e.g. GINI, ...

The most important advantages of decision trees include:

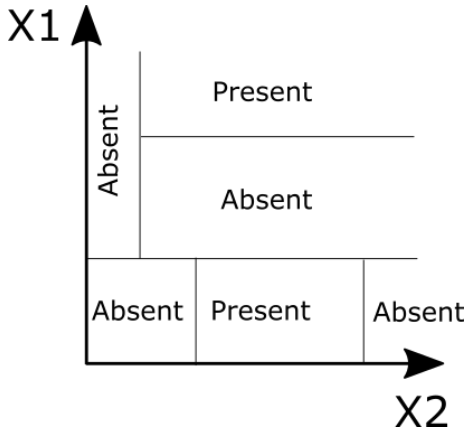

Figure 4.2: Visualization of a classification tree in a two dimensional input space.

- 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

This overview is largely based on Elith, Leathwick, and Hastie 2008 and 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 an 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. Combine the new and old classifiers.

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 overfitting 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 $J \in \{4, \ldots, 8\}$, additionally they observed that the specific value of J in this set has little effect on the performance of the classifier. Hence, we usually use cross validation to obtain optimal values for the learning rate 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).

4.2 Presence only data

Instead of having access to presence-absence data it happens quite often that we only have presence data. This is e.g. the case when we use the records of a natural history museum, herbarium, ... One method of dealing with this kind of data is by extending the methods from Section 4.1, this is done in Section 4.2.1. Another popular approach is Maximum Entropy modeling (Phillips, Anderson, and Schapire 2006; Phillips and Dudík 2008), a short overview of this method is given in Section 4.2.2. Finally we refer to (Pearce and Boyce 2006) for different methods to deal with and additional information on presence-only data.

Finally, 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.

unify MaxEnt / pesudo-absences by introducing IPP models

4.2.1 Classification with pseudo-absences

connection with IPP models

discuss the possibility of using weighted data (e.g. for decision trees etc.) see paper.

Read/find some papers on classification methods with unbalanced data-sets.

discuss effect of using pca on background vs occurrence data

4.2.2 Maximum Entropy modeling

Perhaps the most popular method to create models from presence-only data is by using Maximum Entropy modelling (MaxEnt Phillips, Anderson, and Schapire 2006; Phillips and Dudík 2008). This introduction follows the interpretation introduced by Elith et al. 2011 instead of the original motivation of Phillips, Anderson, and Schapire 2006. This is done mainly because in our opinion it is the most intuitive explanation of the algorithm.

Usually the MaxEnt algorithm is applied to raster data (however this is not necessary). The extent of the raster defines the study are L. Since we usually know the covariate values $\boldsymbol{z}$ of each cell (using e.g. GIS layers) we can define a probability distribution of the covariates $f(\boldsymbol{z})$ which is

implied by defining a uniform distribution over L. We then try to find the distribution $f_1(\boldsymbol{z})$ of the features of the cells where a presence was recorded. This distribution is defined as the distribution which minimizes the Kullback-Leibler divergence with respect to $f(\boldsymbol{z})$ and such that the mean of values of the features (with respect to $f_1 z$) are close to the empircal means of the presence sites.

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). To avoid overfitting a regularization term is added, more particularly a lasso penalty (see Section 5.2.2).

Advantages inlcude:

- MaxEnt has been shown to perform well in a variety of comparative studies (e.g. Elith* et al. 2006).

Some main disadvantages are:

- The output and interpretation is dependent on the scale of the cells.
- MaxEnt is perceived as a black-box model.
- No standard errors are available.
- There are few model checking tools available.

However, most of these problems can be "solved" by utilizing the connection with inhomogeneous poison models (IPP, Fithian and Hastie 2013; Renner and Warton 2013), yet this equivalence is rarely used in practice and as of yet not implemented in the standard MaxEnt software.

4.3 Taking the scale hierarchy into account

In Section 3.3 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 for 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 remote 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.

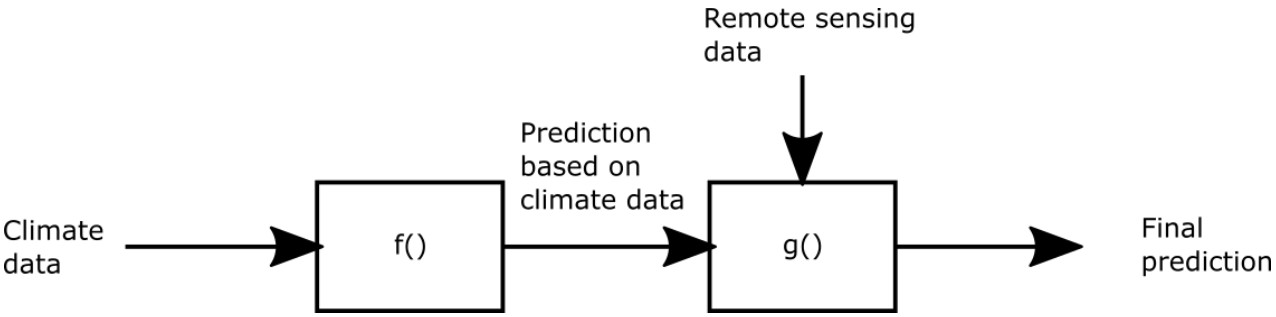

Figure 4.3: Visualization of the hierarchical model.

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 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 further investigated.

Chapter 5

Reducing the number of explanatory variables

The explanatory variables used in SDMs are often correlated. Such high correlation can be an indication of redundant information in the data-set. Moreover, since there is usually no knowledge about which variables make up the niche of a species and irrelevant predictors might be included into the model. 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 5.1 we introduce methods that transform the input space. In Section 5.2 techniques that penalize "large" models are introduced. Finally, Section 5.3 deals with step-wise variable selection methods.

elaborate on problems when using models for prediction vs interpretable models, see Dorman et al. and Harel 2001 (see paper references Dorman)

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.

5.1 Dimensionality reduction

Discuss the effect of the number of background observations ?

Dimensionality reduction techniques can be used to obtain a new, often lower-dimensional,

representation of important structures in the input space. This is often useful because two or more explanatory variables can be a proxy 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).

5.1.1 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 system 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 $\boldsymbol{b}$ is constructed such that

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

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. In this figure there is seems to be an inherent one dimensional subspace (the diagonal line) around which the observations are scattered.

Once we have calculated the principal components we can use them as input for one of the models from Chapter 4. Because the new variables are uncorrelated usually the variance of the estimated coefficients is lower than when using the original variables.

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

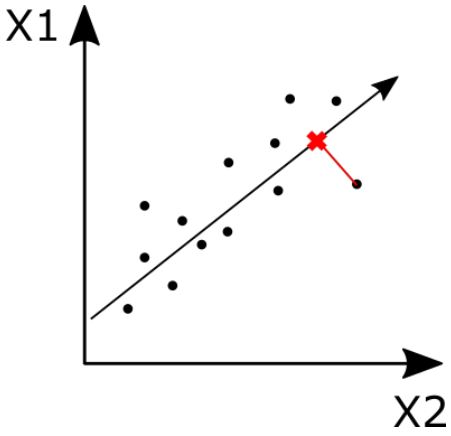

Figure 5.1: Visualization of the a typical scenario where PCA is useful.

cross-validation to select the number of principal components.

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 nonlinear 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.

? mention that we "implicitly" assume that the outcome varies mainly along the PCs with the largest eigenvalues ?

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).

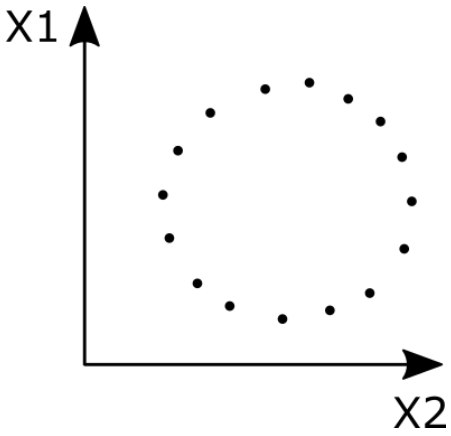

Figure 5.2: Visualization of the a scenario where PCA is useless.

5.1.2 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.

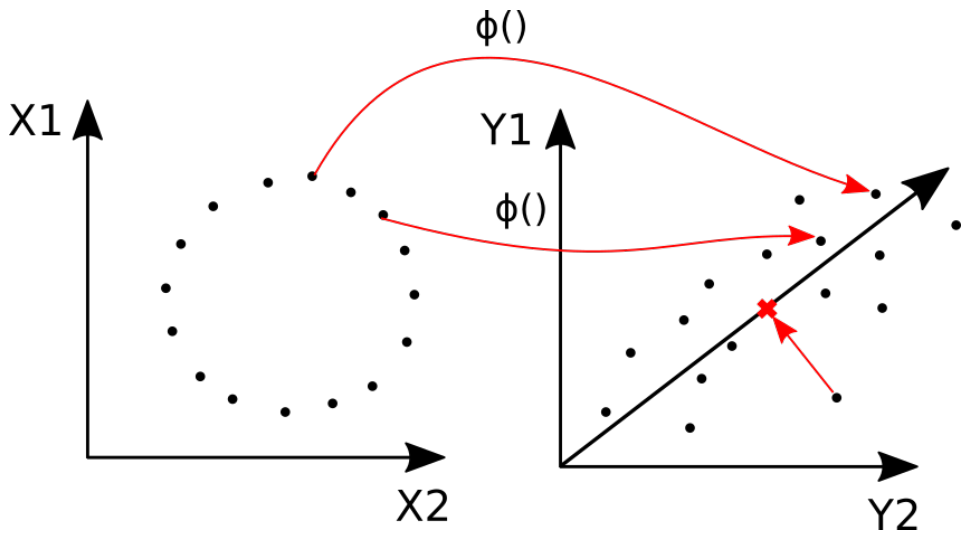

Figure 5.3: Visualization of kernel PCA.

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 $\phi(\cdot)$.

A popular choice is the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}'\boldsymbol{y} + c)^d$, with c some non-zero constant. It can then be shown that the corresponding Hilbert space is the vector space spanned by the products of the vector entries up to degree d. Other choices include the radial basis kernel, $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{\sigma}\right)$, and the sigmoid kernel, $k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(v\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{\sigma})$.

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).

5.1.3 Presence versus background data

5.2 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

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

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 (Friedman et al. 2015) package.

5.2.1 Ridge regression / L_2 regularization

Ridge regression (also called L_2 regularization) is obtained when we use the Euclidean norm of the coefficients as penalization function. Hence, we set

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

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 $\lambda >>$ 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.

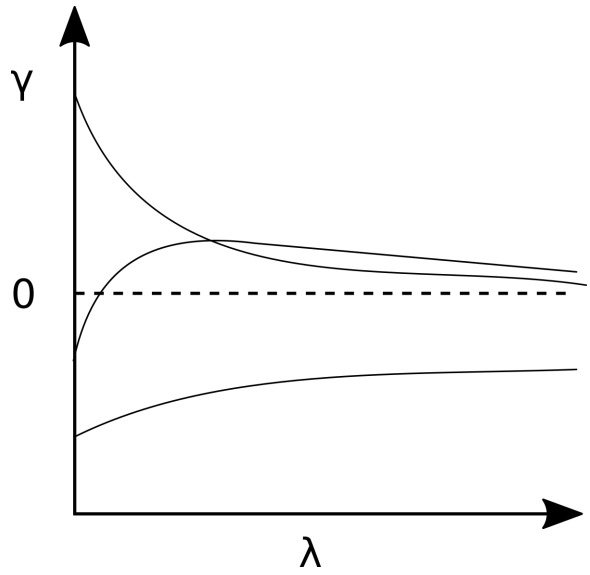

Figure 5.4: A stereotypical evolution of the regression parameters in combination with the ridge parameter.

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.

Disadvantages include:

- Ridge regression keeps all the predictors in the model, i.e. usually no coefficients are equal to zero.
- Ridge regression type estimators are usually biased.

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

5.2.2 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:

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

This L_1 penalty terms is also often 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.

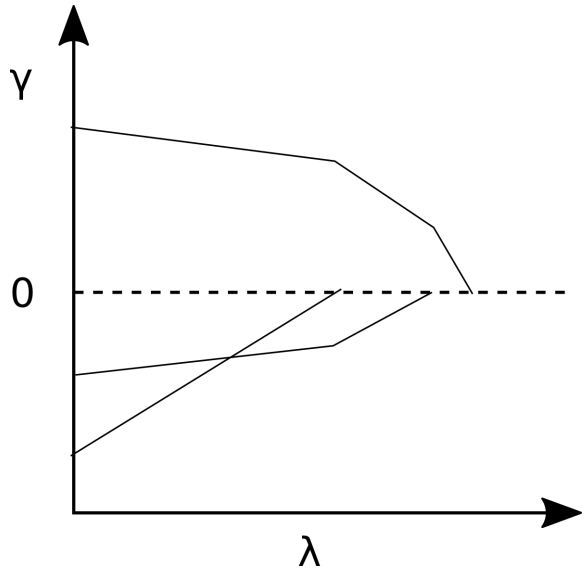

Figure 5.5: Visualization of the typical evulation of the regression parameters in combination with the lasso parameter.

Usually the shrinkage parameter is selected by using cross-validation. 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.

The main advantage is that the lasso performs both shrinkage and parameter selection at once.

5.3 Subset selection methods

Subset selection methods are an older method 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 we will shortly discuss three different techniques.

5.3.1 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

check the number of predictors ...

variables this method is infeasible for our purposes.

5.3.2 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 $p < n$.

Thirdly, also forward-stepwise selection is often used. This method is basically the reverse of the backward-stepwise method. More particularly, we start with the model with only an intercept. Then we add the variable that leads to the largest improvement in the selection criterion. The main advantage of this method over backward-stepwise selection is that it's in general 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 two predictors need to be included at the same time in order to minimize some selection criterion.

5.3.3 Univariate pre-screening

Underlying idea

Univariate pre-screening is closely related to backward-stepwise regression. Just as in backward-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 we 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 applied, 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.

The implementation of this method comes from 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.

5.4 Taking the scale hierarchy into account

double check which kind of models / residuals the used in the paper

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

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

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 throughout this thesis.

5.5 Meaningful combinations of classification and

For each method of Chapter 4 we will consider a "vanilla" model. 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. The vanilla GAM consists of using all predictors together with GCV to select the smoothness parameter. 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 5.1.3 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 background and presence data.
2. a version where the PCA is performed only on the background data.
3. a version where the PCA is performed only presence data.

In total

	PCA	kernel PCA	L_2 penalty	L_1 penalty	subset 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.

[2]Regularization of the "wiggliness" instead of the coefficients, see Section 4.1.2
[3]Using five fold CV to select the L_1 regularization parameter instead of using the default parameter.

Chapter 6

Applications

Chapter 7

Simulation study

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.

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.

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

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. DOI: `10.1073/pnas.0901650106`.

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. DOI: `10.1111/jbi.12225`.

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. DOI: `10.1111/j.1600-0587.2012.07348.x`.

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. DOI: 10.1111/j.1365-2656.2008.01390.x.

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. DOI: 10.1146/annurev.ecolsys.110308.120159.

Elith*, Jane et al. (2006). "Novel methods improve prediction of species' distributions from occurrence data." en. In: *Ecography* 29.2, pp. 129–151. DOI: 10.1111/j.2006.0906-7590.04596.x.

Elith, Jane et al. (2011). "A statistical explanation of MaxEnt for ecologists." en. In: *Diversity and Distributions* 17.1, pp. 43–57. DOI: 10.1111/j.1472-4642.2010.00725.x.

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. DOI: 10.1214/13-AOAS667.

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. DOI: 10.1214/aos/1016218223.

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. DOI: 10.1016/S0304-3800(02)00204-1.

Guisan, Antoine and Wilfried Thuiller (2005). "Predicting species distribution: offering more than simple habitat models." en. In: *Ecology Letters* 8.9, pp. 993–1009. DOI: 10.1111/j.1461-0248.2005.00792.x.

Guisan, Antoine and Niklaus E. Zimmermann (2000). "Predictive habitat distribution models in ecology." In: *Ecological Modelling* 135.2–3, pp. 147–186. DOI: 10.1016/S0304-3800(00)00354-9.

Hastie, Trevor and Werner Stuetzle (1989). "Principal Curves." In: *Journal of the American Statistical Association* 84.406, pp. 502–516. DOI: 10.1080/01621459.1989.10478797.

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. DOI: 10.1111/j.1466-8238.2007.00345.x.

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. DOI: 10.1016/j.ecolmodel.2012.07.028.

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. DOI: 10.1016/0893-6080(89)90020-8.

Karatzoglou, Alexandros et al. (2004). "kernlab – An S4 Package for Kernel Methods in R." In: *Journal of Statistical Software* 11.9, pp. 1–20.

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

Marra, Giampiero and Simon N. Wood (2011). "Practical variable selection for generalized additive models." In: *Computational Statistics & Data Analysis* 55.7, pp. 2372–2387. DOI: 10.1016/j.csda.2011.02.004.

Marx, Brian D. (1996). "Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression." In: *Technometrics* 38.4, pp. 374–381. DOI: 10.2307/1271308.

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. DOI: 10.1111/ecog.00845.

Miller, Jennifer and Janet Franklin (2002). "Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence." In: *Ecological Modelling* 157.2–3, pp. 227–247. DOI: 10.1016/S0304-3800(02)00196-5.

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. DOI: 10.1016/j.ecolmodel.2015.01.019.

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. DOI: 10.1890/07-1289.1.

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. DOI: 10.1111/j.1365-2664.2005.01112.x.

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. DOI: 10.1046/j.1466-822X.2003.00042.x.

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. DOI: 10.1111/j.0906-7590.2004.03740.x.

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. DOI: 10.1016/j.ecolmodel.2005.03.026.

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. DOI: 10.1111/j.0906-7590.2008.5203.x.

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.

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. DOI: 10.1111/j.1541-0420.2012.01824.x.

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. DOI: 10.1038/323533a0.

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. DOI: 10.1111/j.1365-2699.2004.01076.x.

Soberón, Jorge (2007). "Grinnellian and Eltonian niches and geographic distributions of species." en. In: *Ecology Letters* 10.12, pp. 1115–1123. DOI: 10.1111/j.1461-0248.2007.01107.x.

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. DOI: 10.1073/pnas.0901637106.

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. DOI: 10.1046/j.0305-0270.2003.00991.x.

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.

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. DOI: 10.1111/j.1541-0420.2008.01116.x.

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. DOI: 10.1111/j.1365-2656.2006.01141.x.

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. DOI: 10.1073/pnas.0901639106.

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. DOI: 10.7927/H4BP00QC.

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. DOI: 10.1016/S0304-3800(02)00193-X.

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