Species distribution modeling with three-step pseudo–absences

Maialen Iturbide

November 10, 2014

Chapter 1

Introduction

This document provides an introduction to species distribution modeling (SDM) with three–step pseudo–absences.

Species distribution modeling from presence–only data is widely practiced due to the lack of absence data in most of the datasets for predictive modeling. Profile techniques use presence–only data, however, they tend to generate overoptimistic predictions and perform worse than group discrimination approaches, which require both presence and absence data (Elith et al. [2006]; Engler et al. [2004]; Chefaoui and Lobo [2008]). The alternative methodological approach is to use group discrimination techniques relative to the available environment or background samples, also known as pseudo–absences.

One of the most simple methods of generating pseudo–absences is to perform a random selection of the entire study area (Jiang et al. [2014]; Maria Teresa et al. [2014]; Sequeira et al. [2014]). However, it rises the risk of introducing false absences into the model from locations that are suitable for the species. Faced with this problem, several authors employ a presence–only algorithm as a preliminary step to move pseudo–absences away in the environmental space (Zaniewski et al. [2002]; Engler et al. [2004]; Barbet-Massin et al. [2012]; Liu et al. [2013]).

The way of generating pseudo–absences strongly influences the results obtained (Lobo et al. [2010]; Wisz and Guisan [2009]; Barbet-Massin et al. [2012]; Hirzel et al. [2001]), as well as the extent from which background is sampled, a constraint distribution of pseudo–absences around presence locations can lead to misleading models while the opposite, can inflate artificially test statistics and predictions, as well as potentially less informative response variables (Jeremy VanDerWal [2009]).

This document shows an example of a full Species distribution modeling process with pseudo–absences generated in three-steps (TS or TSKM method for pseudo–absence data generation), using for that functions from `mopa` package in R.

If you want to know more about SDM in R, you could consult, for example, documentation from package `dismo` made by Robert J. Hijmans and Jane Elith.

1.1 Species occurrence data

Regarding presence data, Hernandez et al. [2006] suggested that research in environmental niche modeling should focus in broad distribution subunits that are based on distinct genetic linages, in this connection Gonzalez et al. [2011] demonstrated that omission error is reduced when biologically meaningful data is modeled. Thus, functions in the `mopa` package are prepared to run with more than one group of presences at the same time (could be a list of either distribution subunits of a single species or distributions of multiple species), anyway, functions also perform with a single group or species (data frame). In this example we use a data set (list) of two phylogenetic groups of *Quercus* sp in Europe (H11 and H5), available with the `mopa` package .

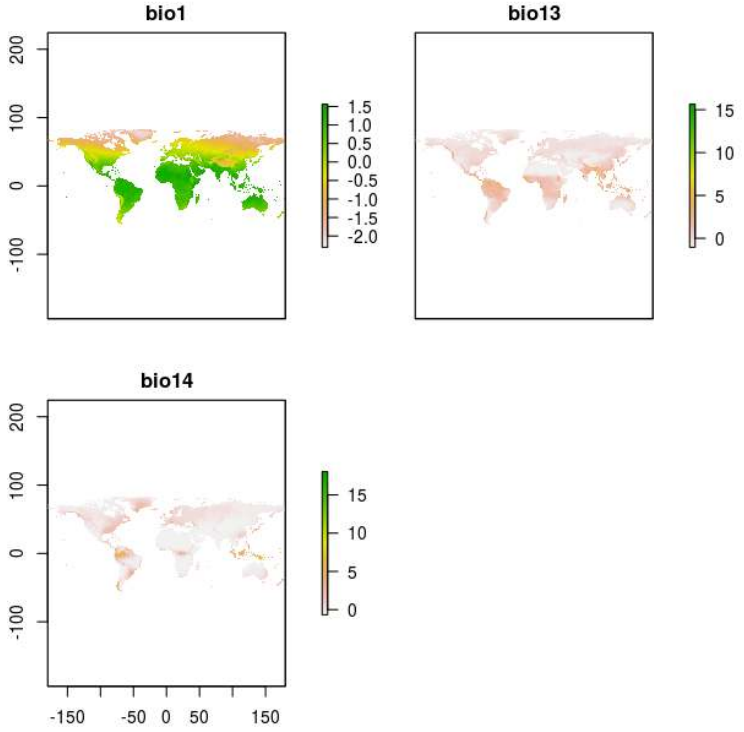
```
> library(mopa)
> data(eu)
> data(Oak_phylo2)
> plot(eu, col="grey")
> for (i in 1:length(Oak_phylo2)){
+   points(Oak_phylo2[[i]], pch="*", cex=0.5,
+          col = colors()[i*50])
+ }
```

1.2 Environmental variables

Predictor variables are typically organized as raster (grid) type files. The set of predictor variables (rasters) can be used to make a '`RasterStack`', which

is a collection of 'RasterLayer' objects (see the raster package for more info) (http://www.cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf)

```
> # RatserStack of envirinmental variables
> data(biostack)
> plot(biostack)
```



Chapter 2

Study area and background

2.1 Creation of the background grid

The regular point grid which covers the continental area can be created as follows:

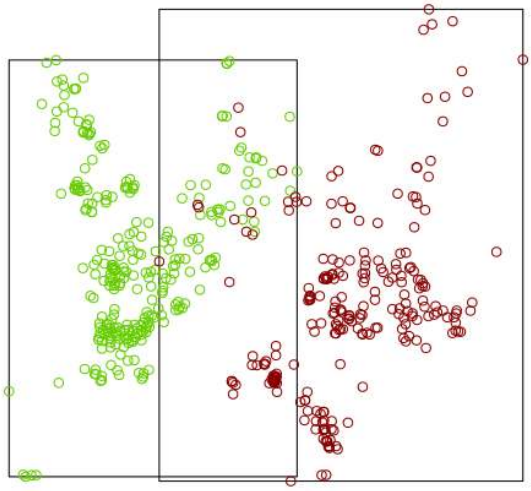
```
> library(raster)
> ac<-xyFromCell(biostack[[1]],  1:ncell(biostack[[1]]))
> ex<-extract(biostack[[1]], ac)
> sp_grid<-SpatialPoints(ac[-which(is.na(ex)),])
> projection(sp_grid)<-CRS("+proj=longlat +init=epsg:4326")
```

2.2 Limit study area to the bounding boxes around presences

```
>  oak.extension<-boundingCoords(Oak_phylo2)
```

Intersection of the background point grid with the bounding boxes. A list with two objects is obtained, (1) bbs: polygon shape of the bounding boxes and (2) bbs.grid: list of data frames of the background point grid limited by the bounding coordinates.

```
> box.grid<-delimit(oak.extension, sp_grid, names(Oak_phylo2))
> plot(box.grid[[1]])
> for (i in 1:length(Oak_phylo2)){
+   points(Oak_phylo2[[i]], col=colors()[i*50])
+ }
```

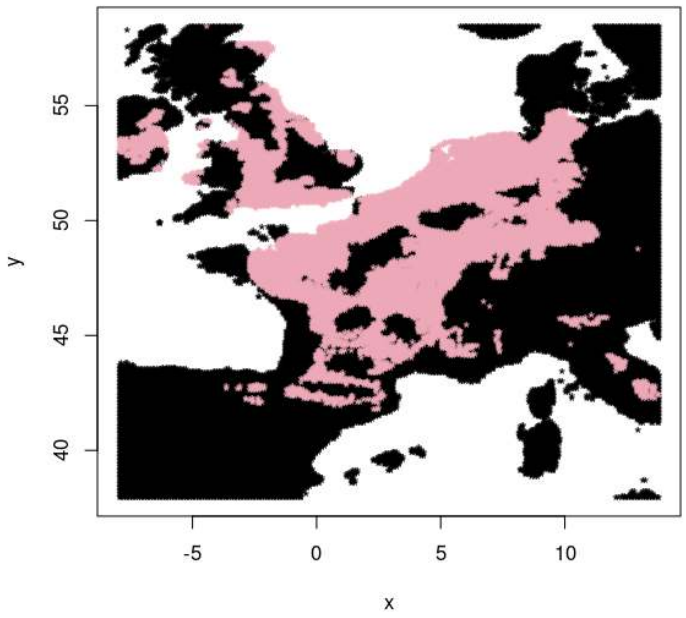


Chapter 3

Three–step pseudo–absences generation

3.1 STEP1: environmental profiling

```
> unsuitable.bg <-OCSVMprofiling(xy = Oak_phylo2,
+                      varstack = biostack,
+                      bbs.grid = box.grid$bbs.grid)
> plot(unsuitable.bg$absence$H11, pch="*")
> points(unsuitable.bg$presence$H11, pch="*", col= "pink2")
```



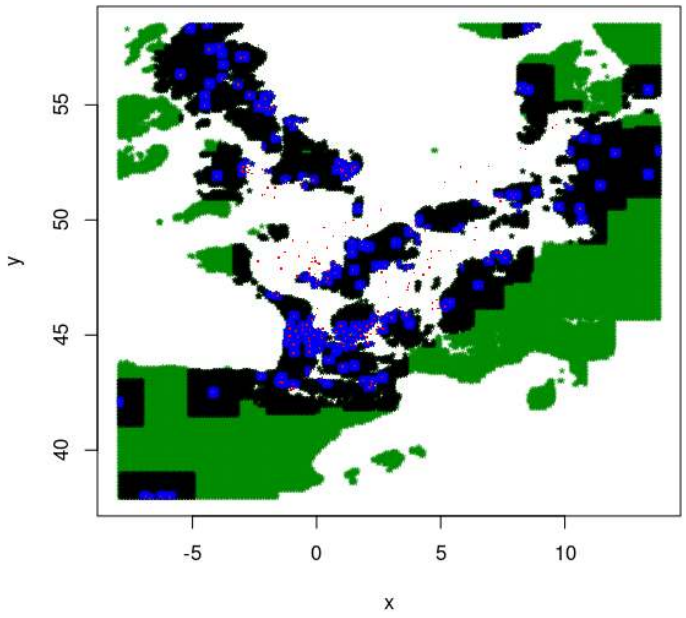
3.2 STEP2: background extents

Sequence of 100 km between distances, from 20 km to the length of the half diagonal of the bounding box:

```
> ext <-bgRadio(xy = Oak_phylo2, bounding.coords = oak.extension,
+         bg.absence = unsuitable.bg$absence, start = 0.166,
+   by = 0.83, unit = "decimal degrees")

[1] "creating background point-grids for species 1 out of 2"
[1] "creating background point-grids for species 2 out of 2"

> plot(ext$H11$km520, col="green4", pch="*")
> points(ext$H11$km120, pch="*")
> points(ext$H11$km20, pch="*", col="blue")
> points(Oak_phylo2$H11, col="red", pch= ".", cex=1.5)
```



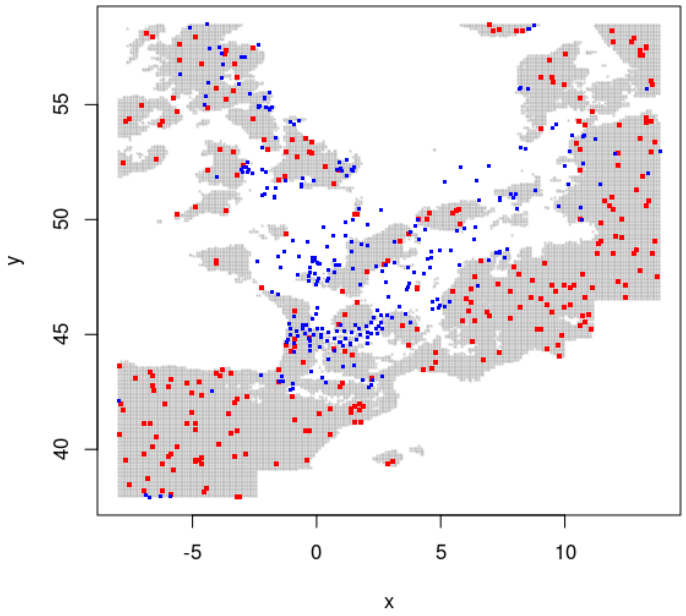
3.3 STEP3: pseudo–absences sampling

3.3.1 At random

```
> pa_random <-PseudoAbsences(xy = Oak_phylo2, bg.grids = ext,
+         exclusion.buffer = 0.0083, prevalence = 0.5,
+   kmeans = FALSE)

[1] "generating pseudo-absences for species 1 out of 2"
[1] "generating pseudo-absences for species 2 out of 2"
```

```
> plot(ext$H11[[5]], pch="*", col= "grey", cex=.5)
> points(pa_random$H11[[5]], col="red", pch=".", cex=4)
> points(Oak_phylo2$H11, col="blue", pch=".", cex=3)
>
```

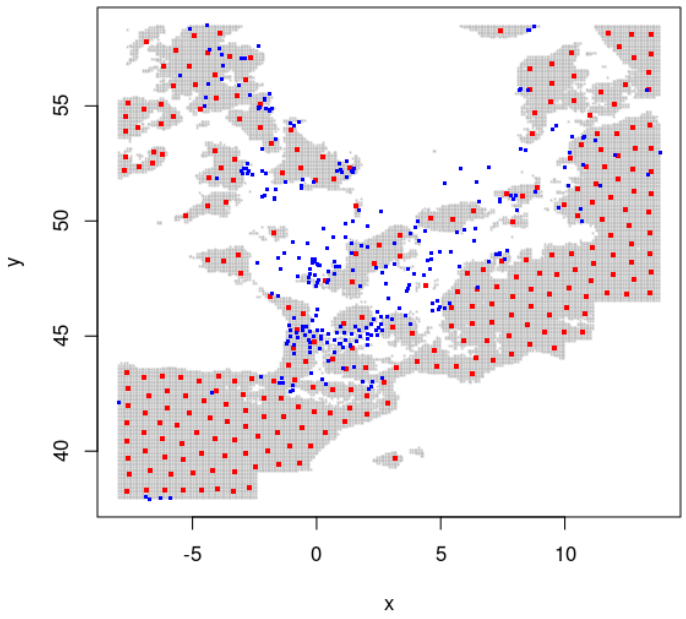


3.3.2 With k–means clustering

```
> pa_kmeans <-PseudoAbsences(xy = Oak_phylo2, bg.grids = ext,
+         exclusion.buffer = 0.0083, prevalence = 0.5,
+   kmeans = TRUE)

[1] "generating pseudo-absences for species 1 out of 2"
[1] "generating pseudo-absences for species 2 out of 2"

> plot(ext$H11[[5]], pch="*", col= "grey", cex=.5)
> points(pa_kmeans$H11[[5]], col="red", pch=".", cex=4)
> points(Oak_phylo2$H11, col="blue", pch=".", cex=3)
```



3.4 Put presences and pseudo–absences together

```
> presaus <-bindPresAbs(presences = Oak_phylo2,
+                       absences = pa_random)
```

Chapter 4

Species distribution modeling

The `allModeling` function.Modelling and cross validation. Algorithms supported are "glm", "svm", "maxent", "mars", "randomForest", "cart.rpart" and "cart.tree"

```
> modirs <-allModeling(data = presaus, varstack = biostack,
+               k = 10, algorithm = "mars", destdir = getwd(),
+               projection = CRS("+proj=longlat +init=epsg:4326"))
```

Named Rdata objects are stored in the specified path. Each Object is given the a name indicating the algorithm, background extent, and species in this order (if a single species is provided no name is given for de species). Character object with listed files is returned. Each Rdata consists of a list with six components:

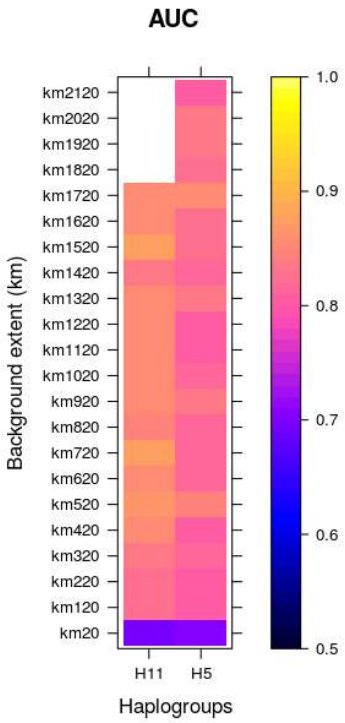
1.-allmod: fitted model with all data for training 2.-auc: AUC statistic in the cross validation 3.-kappa: kappa statistic in the cross validation 4.-tss: true skill statistic in the cross validation 5.-mod: fitted model with partitioned data 6.-p: cross model prediction

To selected the model corresponding to the geographical extent beyond which the AUC scored by the model does not increase we need to load the generated data and extract auc values.

```
> auc_mars <-loadTestValues(data = presaus, test = "auc",
+                           algorithm = "mars")

[1] "loading values for species 1"
[1] "loading values for species 2"

> library(lattice)
> levelplot(auc_mars ,aspect=5 ,at =seq(0.5,1,0.01),
+           col.regions=bpy.colors, xlab="Haplogroups",
+           ylab="Background extent (km)", main = "AUC")
```



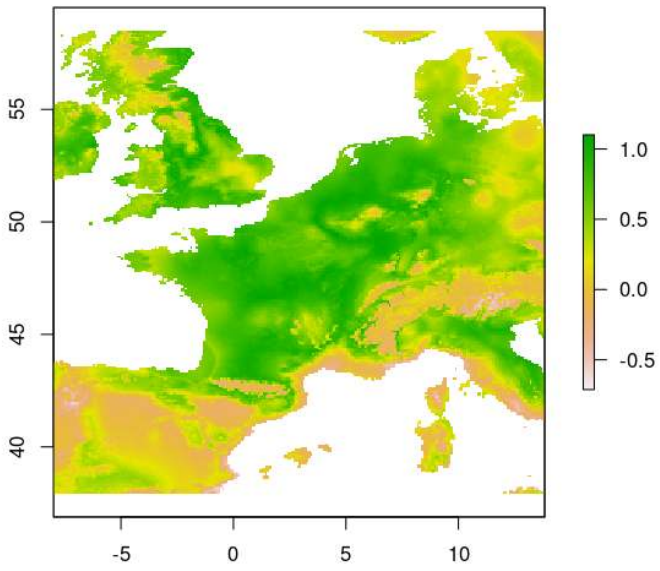
To extract the extent at which maximum AUC values is scored:

```
> ind<-indextent(auc_mars)
> ind

km1220 km1120
    13     12
```

Thus, the `ind` object in this example gives the index of the background extent to be considered for each group/species and is going to be used to extract definitive model components and data. For example:

```
> def <-loadDefinitiveModel(presaus, ind, "allmod", "mars")
> #example of prediction
> projectionland <-biomat(cbind(box.grid[[2]][[1]],
+                     rep(1,nrow(box.grid[[2]][[1]]))),
+                     biostack)
> p <-predict(def[[1]], projectionland[,-1])->p
> spp<-SpatialPixelsDataFrame(box.grid[[2]][[1]],
+                             as.data.frame(p))
> ras<-raster(spp)
> plot(ras)
```


55
50
45
40
-5
0
5
10
1.0
0.5
0.0
-0.5

Bibliography

M.a Barbet-Massin, F.a Jiguet, C.H.b c Albert, and W.c Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 3(2):327–338, 2012. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84867148221&partnerID=40&md5=7e5e3c04a4fd9630c4e6ac6a3783f965. cited By (since 1996)83.

Rosa M. Chefaoui and Jorge M. Lobo. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, 210(4):478–486, February 2008. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2007.08.010. URL http://www.sciencedirect.com/science/article/pii/S030438000700419X.

J. Elith, C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129âĂŞ151, 2006.

Robin Engler, Antoine Guisan, and Luca Rechsteiner. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41(2):263âĂŞ274, 2004. ISSN 1365-2664. doi: 10.1111/j.0021-8901.2004.00881.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.0021-8901.2004.00881.x/abstract.

S.C.a c Gonzalez, J.A.a b Soto-Centeno, and D.L.a Reed. Population distribution models: Species distributions are better modeled using biologically relevant data partitions. *BMC Ecology*, 11, 2011. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-80052849534&partnerID=40&md5=e3e612b2339e279eb9ba7de593016c22. cited By (since 1996)5.

Pilar A. Hernandez, Catherine H. Graham, Lawrence L. Master, and Deborah L. Albert. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5):773–785, October 2006. ISSN 1600-0587. doi: 10.1111/j.0906-7590.2006.04700.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.0906-7590.2006.04700.x/abstract.

A. H. Hirzel, V. Helfer, and F. Metral. Assessing habitat-suitability models with a virtual species. *Ecological modelling*, 145(2):111âĂŞ121, 2001. URL http://www.sciencedirect.com/science/article/pii/S0304380001003969.

Luke P. Shoo Jeremy VanDerWal. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, (4):589–594, 2009. doi: 10.1016/j.ecolmodel.2008.11.010.

Y.a b Jiang, T.c Wang, C.A.J.M.c De Bie, A.K.c Skidmore, X.d Liu, S.a Song, L.e Zhang, J.f Wang, and X.a Shao. Satellite-derived vegetation indices contribute significantly to the prediction of epiphyllous liverworts. *Ecological Indicators*, 38:72–80, 2014. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84888585922&partnerID=40&md5=dabc8cdf6c15fd51d8f84e3aa47a4c3c. cited By (since 1996)0.

C.a Liu, M.a White, G.a Newell, and P.b Griffioen. Species distribution modelling for conservation planning in victoria, australia. *Ecological Modelling*, 249:68–74, 2013. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84872138736&partnerID=40&md5=53f2d771b5ee5f5a33a7ce2eb6a05b6a. cited By (since 1996)3.

Jorge M. Lobo, Alberto JimÃl'nez-Valverde, and JoaquÃ■n Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103âĂŞ114, 2010. ISSN 1600-0587. doi: 10.1111/j.1600-0587.2009.06039.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2009.06039.x/abstract.

C.a Maria Teresa, G.b Antoine, C.b Carmen, S.c Tiziana, L.a Anna, and C.a Maria Laura. A multi-temporal approach to model endangered species distribution in europe. the case of the eurasian otter in italy. *Ecological Modelling*, 274:21–28, 2014. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84890835735&partnerID=40&md5=9a63121595387a5cd6864d48e07fbde7. cited By (since 1996)0.

A.M.M.a Sequeira, C.a b Mellin, D.A.a Fordham, M.G.c Meekan, and C.J.A.a d Bradshaw. Predicting current and future global distributions of whale sharks. *Global Change Biology*, 20(3):778–789, 2014. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84893703315&partnerID=40&md5=d5f92a9819f1daca8ca7e109c475f6f3. cited By (since 1996)0.

Mary S. Wisz and Antoine Guisan. Do pseudo-absence selection strategies influence species distribution models and their predictions? an information-theoretic approach based on simulated data. *BMC Ecology*, 9(1):8, April 2009. ISSN 1472-6785. doi: 10.1186/1472-6785-9-8. URL http://www.biomedcentral.com/1472-6785/9/8/abstract.

A.E. Zaniewski, A. Lehmann, and J.M. Overton. Predicting species spatial distributions using presence-only data: a case study of native new zealand

ferns. *Ecological Modelling*, 157(2):261–280, November 2002. doi: 10.1016/S0304-3800(02)00199-0.