Modelling the potential distribution of Invasive species, Ringed necked parakeet, in the United Kingdom

By Wale Fagbamila

Species distribution modelling (sometimes known as habitat suitability modelling) is the process in which an Algorithm is produced by amalgamating species occurrence data and environmental data which will produce a model predicting the suitability of a location for that species. These models have gained popularity for the various ways in which they can be applied which include determining opportunities to: assess the impact of climate change on native species, to manage and prioritize conservation measures, and predicting species invasions.

To map the habitat suitability of P. krameri I used the Bioclim model which is a profile modelling technique. Bioclim was one of the first species distribution models to be used widely and functions by using presence only data in conjunction with environmental data which will predict the probability of species occurrences in a specified location. (Booth, Nix, Busby and Hutchinson, 2013). A final score between 0-1 is provided, 1 being the maximum probability that the species can occur on the map and 0 being where the species is unlikely to be found.

Workspace organization

To begin with a pair of folders are created to organise the workspace:

dir.create(path = “data”) dir.create(path = “output”)

Install additional R packages

Next, there are six additional R packages that are to be installed:

1.dismo 2.maptools 3.rgdal 4.raster 5.sp

Components of the model

Species distribution models (SDM) uses species occurrence data with environmental data to then project the environmental niche space of a species:

1. Occurrence data: this is the geographic coordinates that a species has been observed. This is known as presence data.
2. Environmental data: these are descriptors of the environment, and includes abiotic and biotic factors.

The getData matrix is given four vital pieces of information:

1.name = “worldclim”: This indicates the name of the data set we would like to download

2.var = “bio”: This tells getData that we want to download all 19 of the bioclimatic variables

3.res = 10: This is the resolution of the data we want to download; in this case, it is 10 minutes of a degree.

4.path = “data/”: This will set the location to which the files are downloaded. In our case, it is the data folder we created at the beginning.

Only the longitude and latitude columns are kept & are rearranged to appear in this order

Duplicate occurrence data is also omitted

```
# Duplicate
obs.data <- NEWROSERINGDATA[,22:23]
obs.data <- obs.data[,c(2,1)]
obs.data <- obs.data[!duplicated(obs.data$decimalLatitude),]
obs.data <- obs.data[!duplicated(obs.data$decimalLongitude),]
```

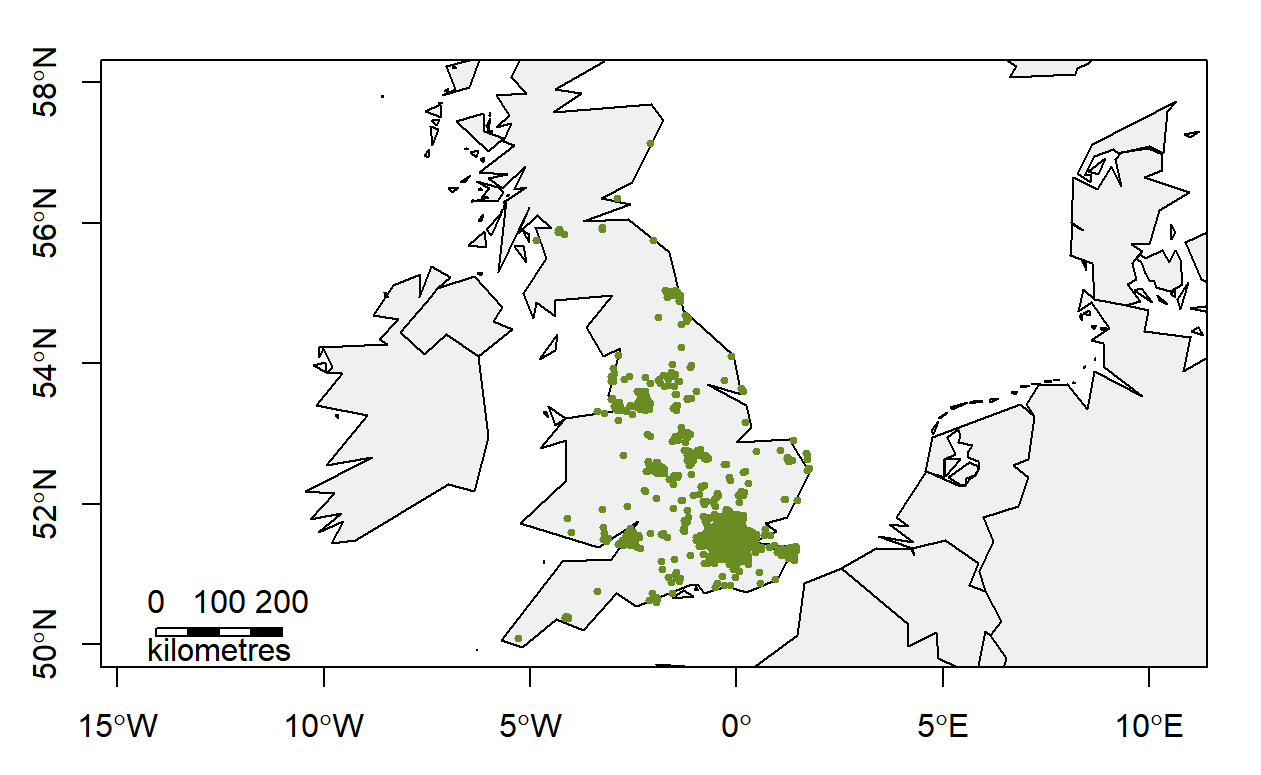
When viewing the obs.data data frame we can see that there are no NA values, so we can now proceed.

To ensure that our species distribution model runs well, it is important to have an idea of how our species of interest is geographically distributed. We find the general latitudinal and longitudinal boundaries and store this information for our code:

```
max.lat <- ceiling(max(obs.data$decimalLatitude))
min.lat <- floor(min(obs.data$decimalLatitude))
max.lon <- ceiling(max(obs.data$decimalLongitude))
min.lon <- floor(min(obs.data$decimalLongitude))
geographic.extent <- extent(x = c(min.lon, max.lon, min.lat, max.lat))
```

It is good practice when coding to run a reality check on your occurrence data by plotting the points on a map.

```
# Load the data to use for our base map
data(wrld_simpl)

# Plot the base map
plot(wrld_simpl,
     xlim = c(min.lon, max.lon),
     ylim = c(min.lat, max.lat),
     axes = TRUE,
     col = "grey95")

# Add the points for individual observation
points(x = obs.data$decimalLongitude,
       y = obs.data$decimalLatitude,
       col = "olivedrab",
       pch = 20,
       cex = 0.75)
# And draw a little box around the graph
box() + scalebar(200, type = "bar", divs = 4, below = "kilometres")
```




```
## integer(0)
```

Building a model and visualizing results

Our occurrence data looks good, so now we can use the bioclimatic variables to create a model. We restrict the biolimatic variable data to the geographic extent of our occurrence data as we are not looking to model the Rose ringed parakeet habitat suitablity globally, but to our area of interest (The United Kingdom).

```
# Crop bioclim data to geographic extent of saguaro
bioclim.data <- crop(x = bioclim.data, y = geographic.extent)

# Build species distribution model
bc.model <- bioclim(x = bioclim.data, p = obs.data)
```

One more step that needs to take before we plot the model on a map is that We generate an object that has the model’s probability of occurrence for the Rose ringed parakeet. We retrieve the predict model from the dismo package:

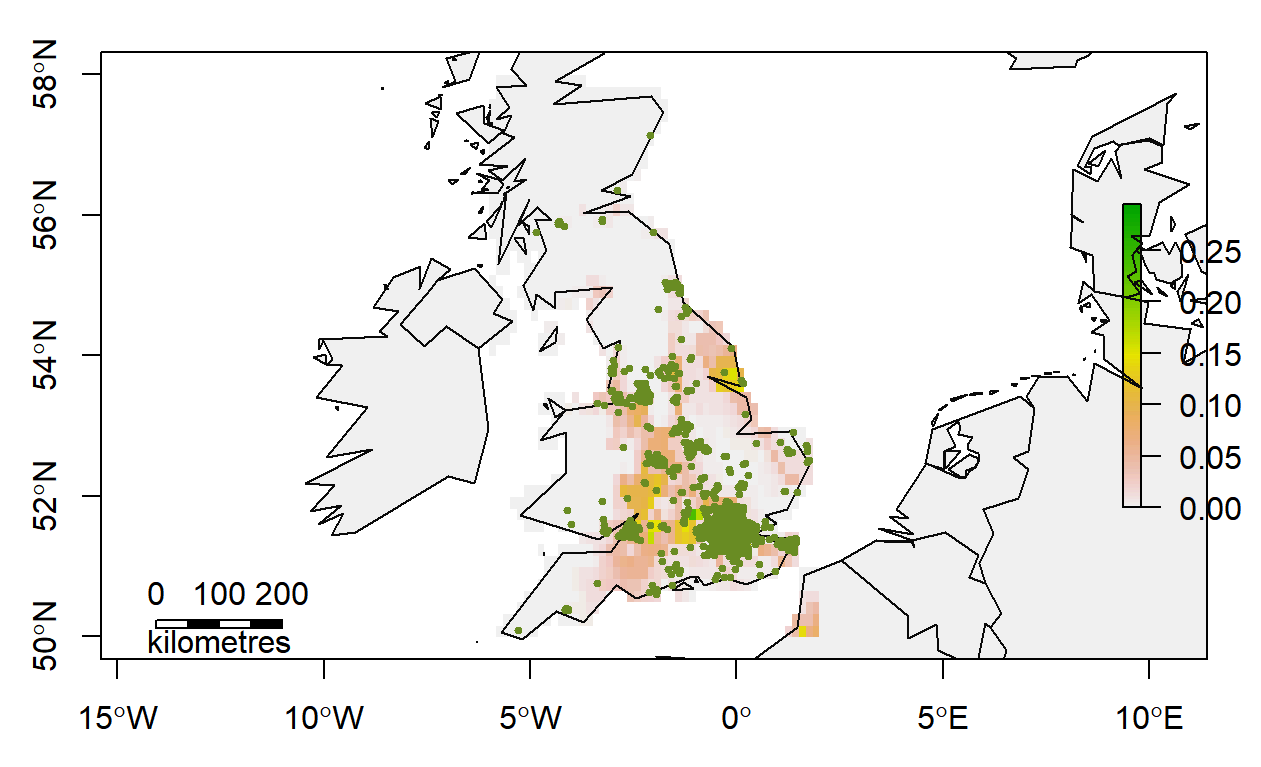
```
# Predict presence from model
predict.presence <- dismo::predict(object = bc.model,
                                   x = bioclim.data,
                                   ext = geographic.extent)
```

Now that we have completed those steps we can now plot and we start by adding a blank map, then we add the model probabilities along with the original observations.

```
plot(wrld_simpl,
     xlim = c(min.lon, max.lon),
     ylim = c(min.lat, max.lat),
     axes = TRUE,
     col = "grey95")

# Add model probabilities
plot(predict.presence, add = TRUE)

# Redraw those country borders
plot(wrld_simpl, add = TRUE, border = "grey5")

# Add original observations
points(obs.data$decimalLongitude, obs.data$decimalLatitude, col = "olivedrab", pch = 20, cex = 0.75) + box()
```

```
## integer(0)
```

```
 scalebar(200, type = "bar", divs = 4, below = "kilometres")
```



This plot shows the probability of occurrence of Rose ringed parakeet across the UK. The maximum probability anywhere on the map is 0.30. This model however, lacks in comparison to its newer counterparts as Bioclim can only factor in continuous environmental predictors only and doesn’t account for the interactions between these environmental variables. (GU eResearch, 2021) Additionally, sampling bias exists greatly in presence only data models, especially in cases where there is an inadequate number of species records available (for example when studying endangered species).

If we want our map to better reflect this, we will need to include background points. Whereas the issue is that we only have presence data for the rose ringed parakeets.

The pseudo-absence point

This model utilises randomly sampled background points (also called pseudo-absence) from a geographic area and effectively they are placed at locations where the focal species is absent and the purpose of this is to avoid bias. Research conducted by Barbet-Massin, Jiguet, Albert and Thuiller (2012) concluded that the most effective species distribution models required data both species presence data and pseudo-absence data. In comparison to profile models, the multiple linear regression model can handle both continuous and categorical models and factors in the interactions between those variables.

```
# Use the bioclim data files for sampling resolution
bil.files <- list.files(path = "data/wc2-5",
                        pattern = "*.bil$",
                        full.names = TRUE)

# We only need one file, so use the first one in the list of .bil files
mask <- raster(bil.files[1])

# Set the seed for the random-number generator to ensure results are similar
set.seed(4648)

# Randomly sample points (same number as our observed points)
background <- randomPoints(mask = mask,     # Provides resolution of sampling points
                   n = nrow(obs.data),      # Number of random points
                   ext = geographic.extent, # Spatially restricts sampling
                   extf = 1.25)             # Expands sampling a little bit
```

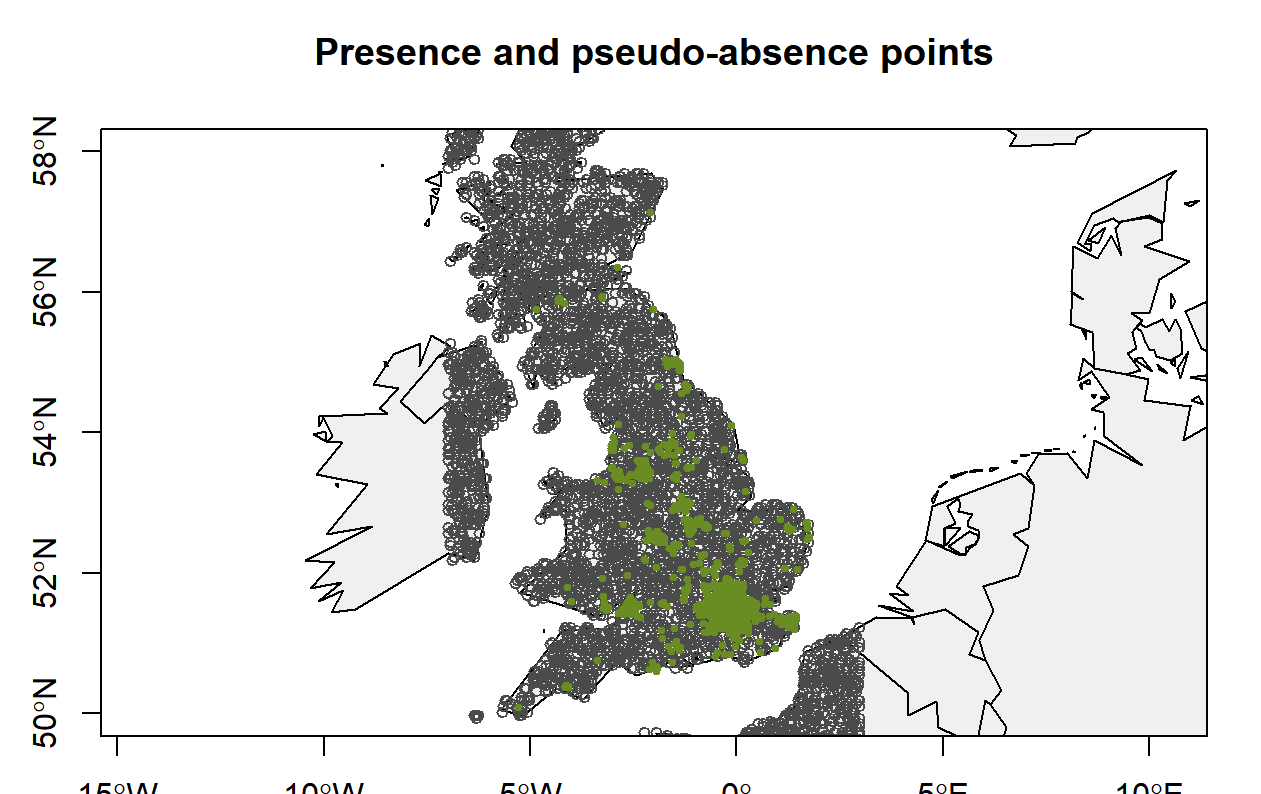
Take a quick look at the background object we just created:

```
head(background)
```

```
##              x        y
## [1,] -1.979167 51.72917
## [2,] -3.437500 50.68750
## [3,]  0.437500 51.18750
## [4,] -3.229167 54.93750
## [5,] -5.312500 56.60417
## [6,] -3.395833 53.02083
```

We can also visualize them on a map, like we did for the observed points:

```
# Plot the base map
plot(wrld_simpl,
     xlim = c(min.lon, max.lon),
     ylim = c(min.lat, max.lat),
     axes = TRUE,
     col = "grey95",
     main = "Presence and pseudo-absence points")

# Add the background points
points(background, col = "grey30", pch = 1, cex = 0.75)

# Add the observations
points(x = obs.data$decimalLongitude,
       y = obs.data$decimalLatitude,
       col = "olivedrab",
       pch = 20,
       cex = 0.75)

box()
```



Now that we have our pseudo-absence points, a post hoc evaluation of the pseudo-absence points model is also conducted. To do this evaluation, I split my model into two segments, the training data and the testing data. 20% of the data was reserved for testing and I utilised the kfold function within the dismo package of R studio was used in order to consistently distribute the observations to random groups. Also, it should be considered that with random sampling to generate the pseudo-absence points, there is the possibility for variation in the predicted range if the code is to be repeated several times.

```
# Arbitrarily assign group 1 as the testing data group
testing.group <- 1

# Create vector of group memberships
group.presence <- kfold(x = obs.data, k = 5) # kfold is in dismo package
```

```
# Separate observations into training and testing groups
presence.train <- obs.data[group.presence != testing.group, ]
presence.test <- obs.data[group.presence == testing.group, ]

# Repeat the process for pseudo-absence points
group.background <- kfold(x = background, k = 5)
background.train <- background[group.background != testing.group, ]
background.test <- background[group.background == testing.group, ]
```

Training and testing the model

Now that we have (1) our pseudo-absence points and (2) separate training and testing data, we can re-build the model, evaluate its performance, and draw a more aesthetically pleasing map. We build the model with the bioclim function as before, but instead of using all the observations in obs.data we only use the training data stored in presence.train:

```
# Build a model using training data
bc.model <- bioclim(x = bioclim.data, p = presence.train)

# Predict presence from model (same as previously, but with the update model)
predict.presence <- dismo::predict(object = bc.model,
                                   x = bioclim.data,
                                   ext = geographic.extent)
```

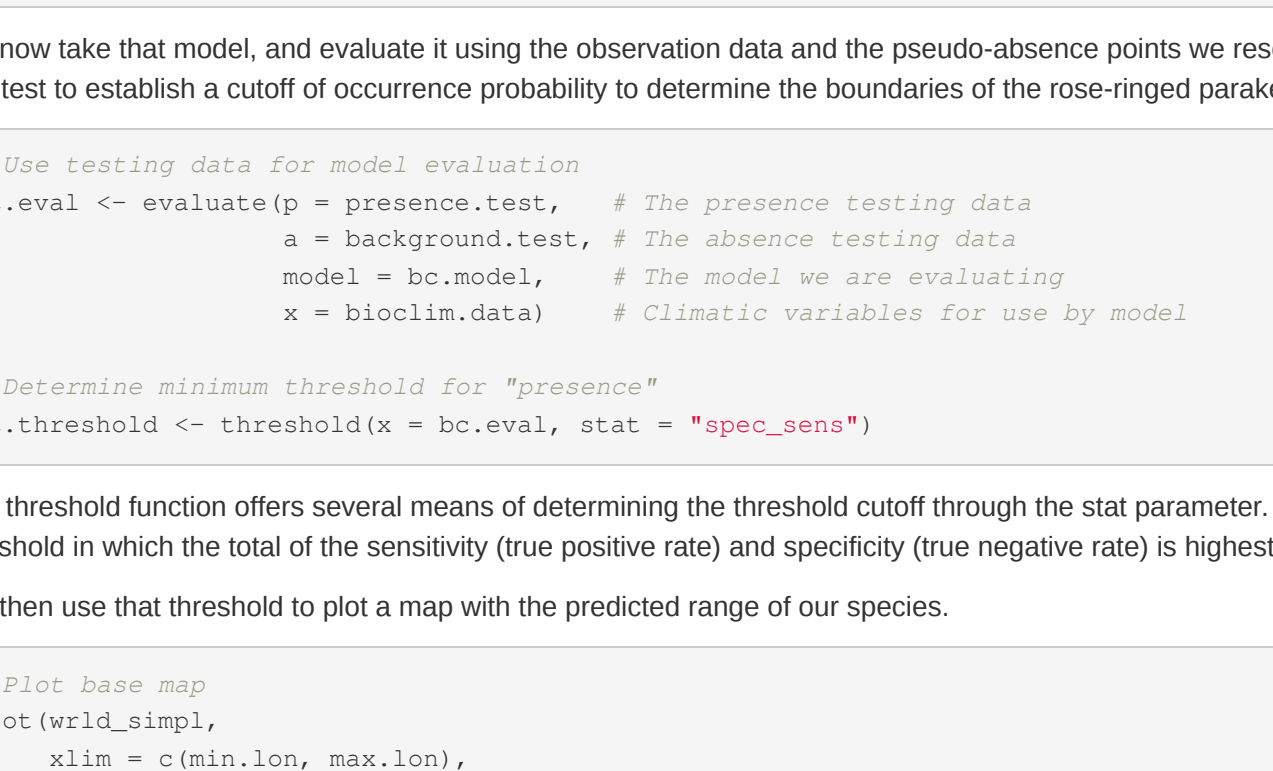
We now take that model, and evaluate it using the observation data and the pseudo-absence points we reserved for model testing. We then use this test to establish a cutoff of occurrence probability to determine the boundaries of the rose-ringed parakeet.

```
# Use testing data for model evaluation
bc.eval <- evaluate(p = presence.test,   # The presence testing data
                    a = background.test, # The absence testing data
                    model = bc.model,    # The model we are evaluating
                    x = bioclim.data)    # Climatic variables for use by model

# Determine minimum threshold for "presence"
bc.threshold <- threshold(x = bc.eval, stat = "spec_sens")
```

The threshold function offers several means of determining the threshold cutoff through the stat parameter. The “spec_sens” function sets the threshold in which the total of the sensitivity (true positive rate) and specificity (true negative rate) is highest.

We then use that threshold to plot a map with the predicted range of our species.

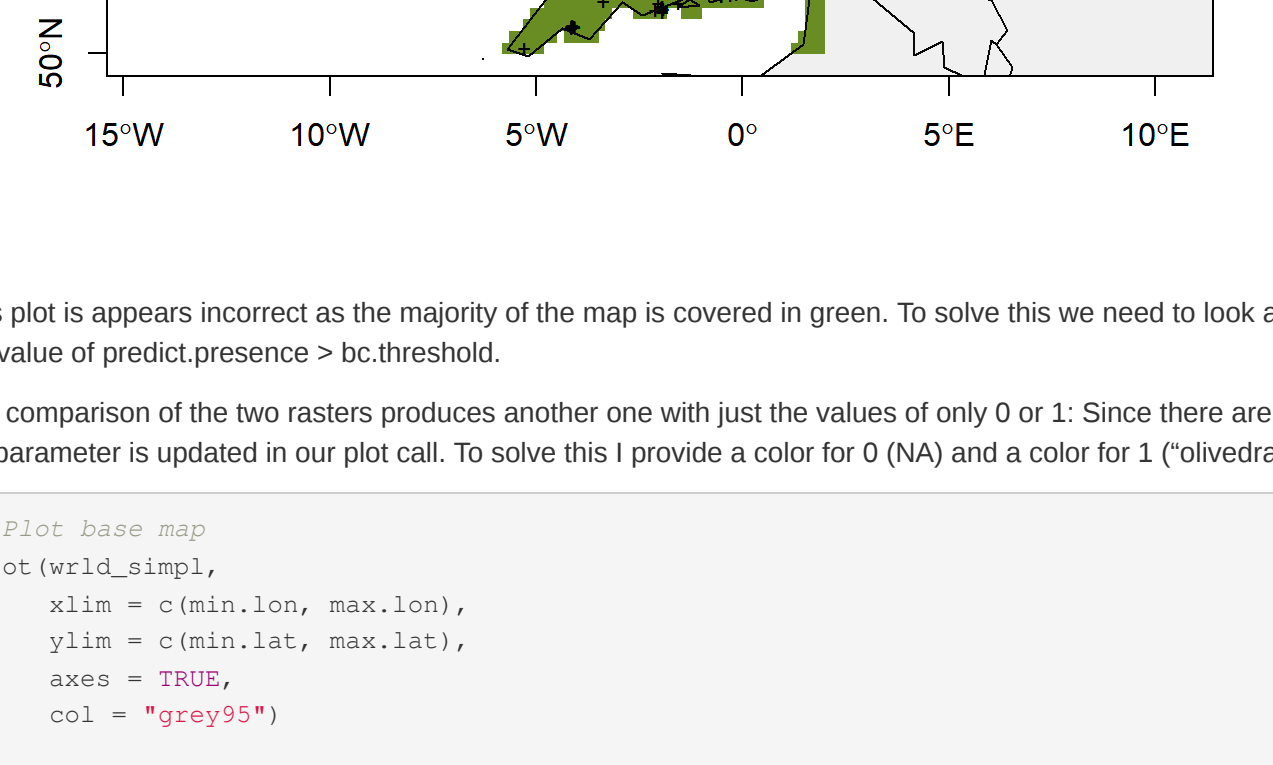
```
# Plot base map
plot(wrld_simpl,
     xlim = c(min.lon, max.lon),
     ylim = c(min.lat, max.lat),
     axes = TRUE,
     col = "grey95")

# Only plot areas where probability of occurrence is greater than the threshold
plot(predict.presence > bc.threshold,
     add = TRUE,
     legend = FALSE,
     col = "olivedrab")

# And add those observations
points(x = obs.data$decimalLongitude,
       y = obs.data$decimalLatitude,
       col = "black",
       pch = "+",
       cex = 0.75)

# Redraw those country borders
plot(wrld_simpl, add = TRUE, border = "grey5")
box()
```



This plot is appears incorrect as the majority of the map is covered in green. To solve this we need to look at what we asked R to plot, so we plot the value of predict.presence > bc.threshold.

The comparison of the two rasters produces another one with just the values of only 0 or 1: Since there are only two values in this comparison , the col parameter is updated in our plot call. To solve this I provide a color for 0 (NA) and a color for 1 (“olivedrab”):

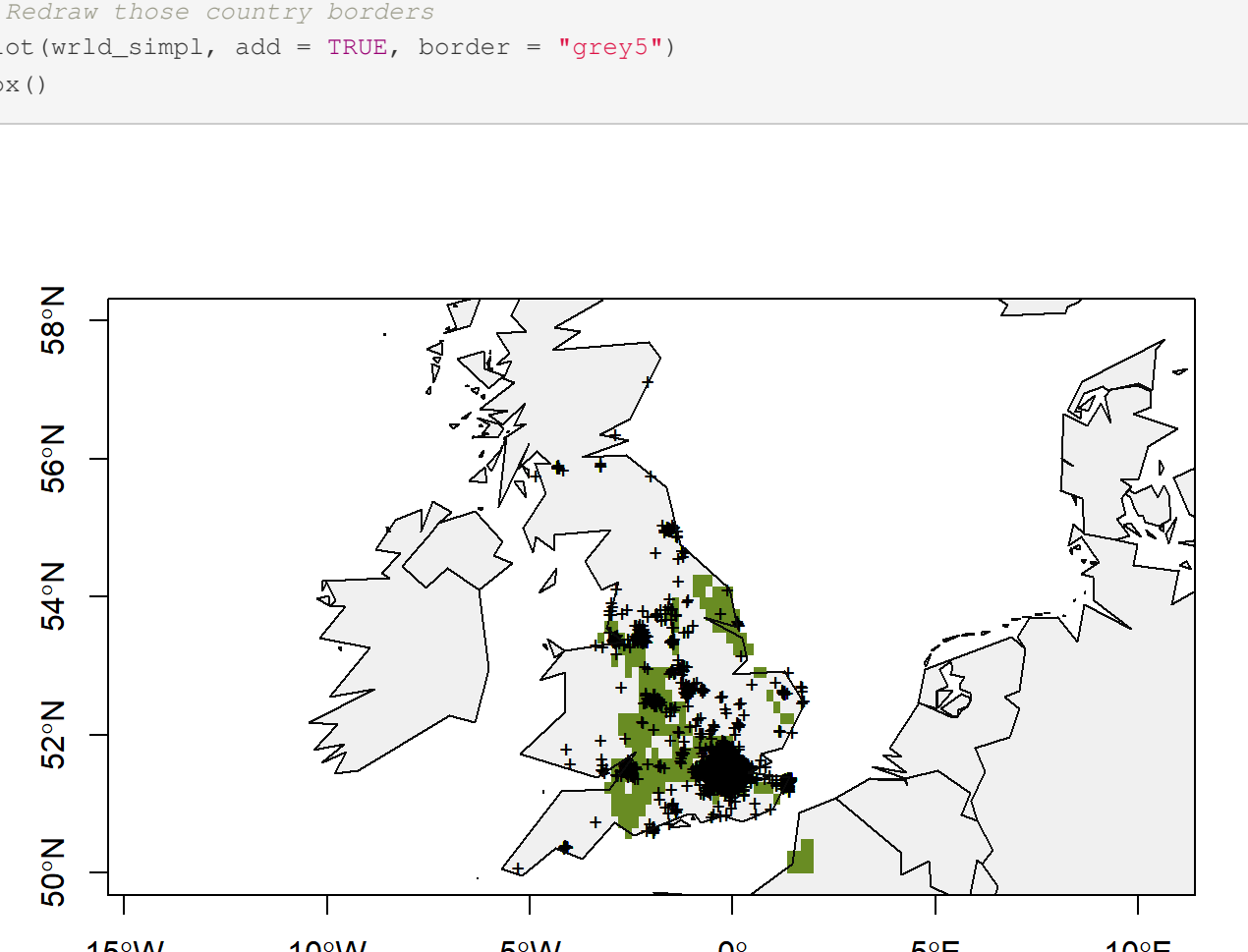
```
# Plot base map
plot(wrld_simpl,
     xlim = c(min.lon, max.lon),
     ylim = c(min.lat, max.lat),
     axes = TRUE,
     col = "grey95")

# Only plot areas where probability of occurrence is greater than the threshold
plot(predict.presence > bc.threshold,
     add = TRUE,
     legend = FALSE,
     col = c(NA, "olivedrab"))

# And add those observations
points(x = obs.data$decimalLongitude,
       y = obs.data$decimalLatitude,
       col = "black",
       pch = "+",
       cex = 0.75)

# Redraw those country borders
plot(wrld_simpl, add = TRUE, border = "grey5")
box()
```



The habitat suitability model produced using the Psuedo-Absence modelling technique which produced a map that presents a categorical classification of the point on the landscape will be suitable or not for the Rose ring parakeet.

Statistics

```
plot(bc.eval,'ROC') # STATS FOR MODEL
```



The AUC (area under the receiver operator curve) is the statistic that is used for assessing the discriminatory capacity of species distribution models. (Jiménez-Valverde, 2011). An AUC of 0.5 or lower suggests that there is no discrimination within our model, a reading of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is interpreted to be outstanding. (Mandrekar, 2010)

Conclusion

In conclusion, this research project established that species distribution modelling can be an effective tool for identifying the suitable habitats for Psittacula krameria. The model provides visual representation and statistical values which enables ecologist and conservationist to make decisions aided by the knowledge that they will gain from spotting areas that are likely to be appropriate for the invasive species to thrive in.