BIO 423 - Lab 4

Benjamin Blonder

Spring 2019

Learning outcomes

R goals: - plot and extract value from raster datasets - use `apply` family of functions - use basic significance tests

Content goals: - Contrast resource-use niches & evaluate hypotheses for different processes underlying patterns - Build environmental niche models

Resource niches

We will begin by revisiting MacArthur's (1958) insectivorous warbler dataset, in which the niches of five species of birds were measured. Species included the Cape May, Yellow-rumped, Black-throated Green, Blackburnian, and Bay-breasted warbler. Data are copied frm Gotelli's `EcoSimR` package.

We will re-analyze MacArthur's original data to determine whether these species indeed have the same or different niches. Recall that the data includes information on the feeding position of each species on diffferent parts of a tree.

Recall the parts of the tree are defined in Figure 2 of MacArthur (1958) as two-part codes:

Part one: 1 = top of tree 6 = base of tree

Part two: B = base of branch M = middle of branch T = tip of branch

In R, these will be coded as column names starting with X, e.g. `X1B`.

```
library(ggplot2) # make sure this library is installed!
library(reshape2)

data_warblers = read.csv(file="dataMacWarb.csv")
# verify the types of variables in the data frame
str(data_warblers)
```

```
## 'data.frame':    5 obs. of  17 variables:
##  $ Species: Factor w/ 5 levels "Bay-breasted warbler",..: 4 5 2 3 1
##  $ X1T    : num  49.9 6.6 12.1 34.8 3.5
##  $ X1M    : num  13.2 4.1 5.7 10.5 1.9
##  $ X1B    : num  0 0.3 0 3.2 1.4
##  $ X2T    : num  20.6 7.8 17.3 15.1 6.5
##  $ X2M    : num  8.3 4.9 8.8 8.3 8
##  $ X2B    : num  0 1.3 0.7 2.7 5.9
##  $ X3T    : num  4 9.3 21.8 13.1 11.4
##  $ X3M    : num  0.5 9.8 14.1 11 19.1
##  $ X3B    : num  0 3.6 1.5 0.7 13.1
##  $ X4T    : num  0 1.7 6.2 0.3 7.7
##  $ X4M    : num  0 1.3 4.7 0.3 8.8
##  $ X4B    : num  0 5.1 4.5 0 10.1
##  $ X5T    : num  0 0 1.4 0 0
```

```
##  $ X5M    : num  0 0.6 0.3 0 0.1
##  $ X5B    : num  0 15.9 0.9 0 2.5
##  $ X6     : num  3.5 27.7 0 0 0
```

```
# verify the names of the columns
names(data_warblers)
```

```
##  [1] "Species" "X1T"     "X1M"     "X1B"     "X2T"     "X2M"     "X2B"
##  [8] "X3T"     "X3M"     "X3B"     "X4T"     "X4M"     "X4B"     "X5T"
## [15] "X5M"     "X5B"     "X6"
```

We can also summarize time spent in different parts of the tree. To do so,

```
# make a new object including all the time spent along one part of the tree
timePosition1 = data.frame(Time.1T=data_warblers$X1T,
                           Time.1M=data_warblers$X1M,
                           Time.1B=data_warblers$X1B)
str(timePosition1)
```

```
## 'data.frame':    5 obs. of  3 variables:
##  $ Time.1T: num  49.9 6.6 12.1 34.8 3.5
##  $ Time.1M: num  13.2 4.1 5.7 10.5 1.9
##  $ Time.1B: num  0 0.3 0 3.2 1.4
```

```
# make a new column in the original data for
# the total % of time spent along any of the '1' positions
?rowMeans
data_warblers$MeanPercentPosition1 = rowMeans(timePosition1)
# the second line does an equivalent calculation - it applies the 'mean' function
# across all rows (MARGIN=1) in the input X
data_warblers$MeanPercentPosition1 = apply(X=timePosition1, MARGIN=1, FUN=mean)
# thus we can also calculate total time in position 1...
data_warblers$SumPercentPosition1 = apply(X=timePosition1, MARGIN=1, FUN=sum)
```

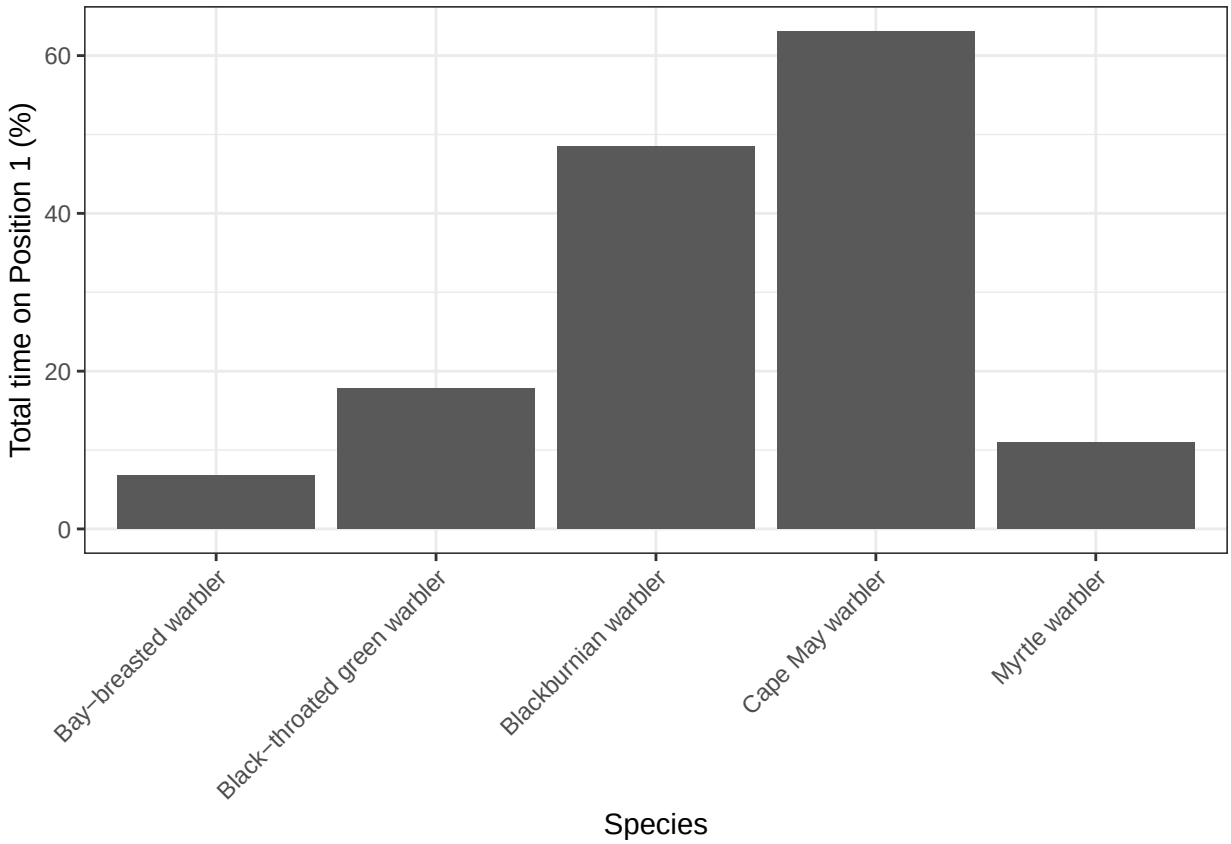
```
# examine mean values for all species
data_warblers[,c("Species","MeanPercentPosition1")]
```

```
##                       Species MeanPercentPosition1
## 1             Cape May warbler            21.033333
## 2               Myrtle warbler             3.666667
## 3 Black-throated green warbler             5.933333
## 4         Blackburnian warbler            16.166667
## 5         Bay-breasted warbler             2.266667
```

```
# extract values for just one species
data_warblers[data_warblers$Species=="Blackburnian warbler",]
```

```
##               Species  X1T  X1M X1B  X2T X2M X2B  X3T X3M X3B X4T X4M X4B
## 4 Blackburnian warbler 34.8 10.5 3.2 15.1 8.3 2.7 13.1  11 0.7 0.3 0.3   0
##   X5T X5M X5B X6 MeanPercentPosition1 SumPercentPosition1
## 4   0   0   0  0             16.16667                48.5
```

```
# plot the usage along the X1T axis of each species
ggplot(data_warblers,aes(x=Species,y=SumPercentPosition1)) +
  geom_bar(stat="identity") +
  theme_bw() +
  ylab("Total time on Position 1 (%)") + xlab("Species") +
  theme(axis.text.x=element_text(angle=45,hjust=1))
```



Next, we will examine how similarly species use different portions of the tree. To do so, we will calculate the Pearson correlation (`https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`) between all of the different resource use positions for each pair of species. For vectors X and Y, the correlation ρ is defined as

$$\rho = \frac{E[(X - \mu(X))(Y - \mu(Y))]}{\sigma(X)\sigma(Y)}$$

where E indicates an expected value, μ a mean, and σ a standard deviation.

Values of ρ closer to 1 indicate more similarity; closer to 0, randomness, closer to -1, more differentiation. If we start with a `m` x `n` matrix of `m` species with `n` variables for each, we will end up with a `m` x `m` matrix of similarities between species.

To calculate a similarity matrix, we first need to reshape the data to only contain numeric values (i.e. losing the `Species` column and transpose it so that each species is a column rather than row (see help for the `cor` function to understand why).

```
# calculate Pearson correlation between all time uses
# (after transposing row/column order to satisfy the
# assumptions of the cor() function
?cor
data_transposed = as.matrix(t(data_warblers[,2:17]))
# verify these column indices get all the data you want
str(data_transposed)
```

```
##  num [1:16, 1:5] 49.9 13.2 0 20.6 8.3 0 4 0.5 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:16] "X1T" "X1M" "X1B" "X2T" ...
##   ..$ : NULL
```

```
# re-assign the names
dimnames(data_transposed)[[2]] =  data_warblers$Species
str(data_transposed)
```

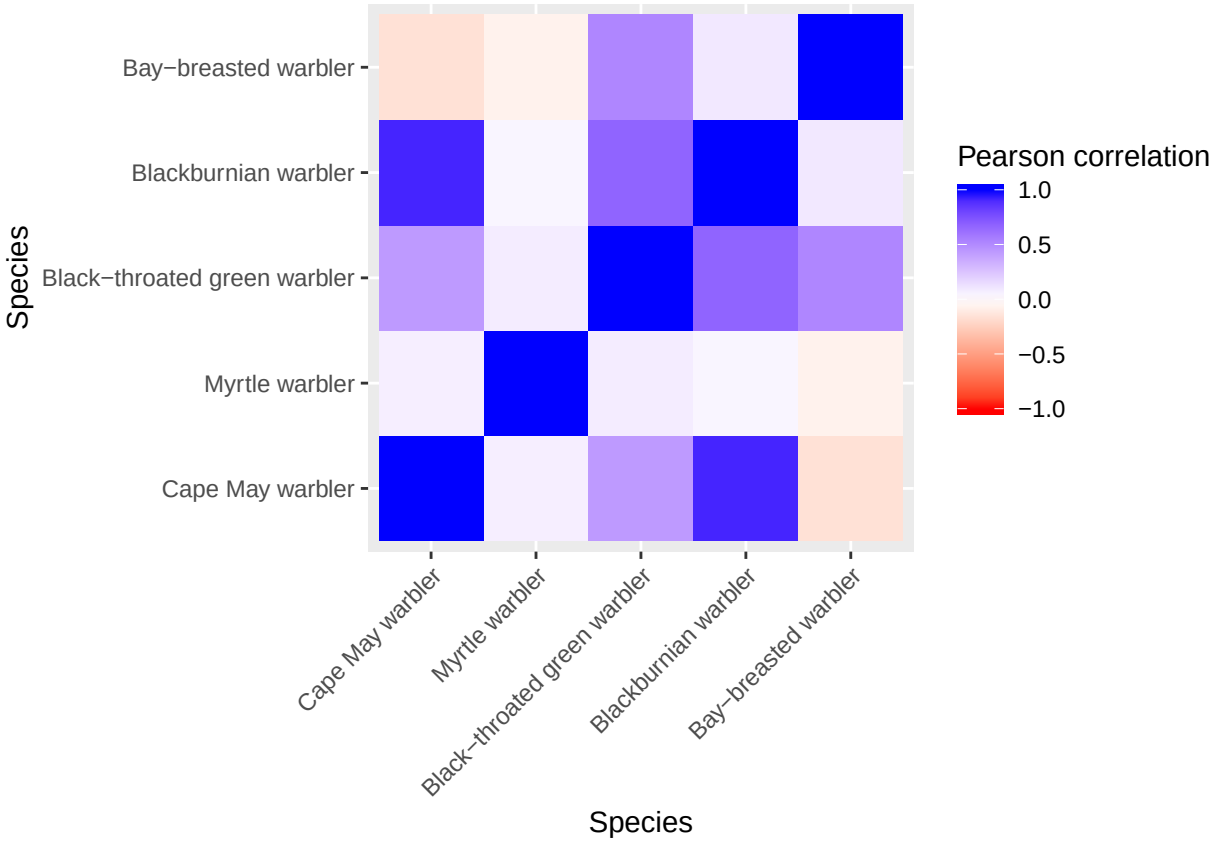
```
##  num [1:16, 1:5] 49.9 13.2 0 20.6 8.3 0 4 0.5 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:16] "X1T" "X1M" "X1B" "X2T" ...
##   ..$ : chr [1:5] "Cape May warbler" "Myrtle warbler" "Black-throated green warbler" "Blackburnian wa
```

```
# calculate pairwise Pearson correlation between all rows (i.e. species)
warbler_similarity = cor(data_transposed)
# note that the correlation matrix is symmetric -
# the similarity between species A&B is the same as between B&A

# reshape the similarity matrix to 'long' format for easy plotting
warbler_similarity_melted = melt(warbler_similarity)
str(warbler_similarity_melted)
```

```
## 'data.frame':    25 obs. of  3 variables:
##  $ Var1 : Factor w/ 5 levels "Cape May warbler",..: 1 2 3 4 5 1 2 3 4 5 ...
##  $ Var2 : Factor w/ 5 levels "Cape May warbler",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ value: num  1 0.0705 0.4342 0.9237 -0.159 ...
```

```
ggplot(warbler_similarity_melted, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + xlab("Species") + ylab("Species") +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white",
   midpoint = 0, limit = c(-1,1), name="Pearson correlation") +
  theme(axis.text.x=element_text(angle=45,hjust=1))
```



Questions

1. Which species spends the most total time at the base of the tree (Position 6)?
2. Which species has the most variable resource use, i.e. has the highest variation in percent time across locations? Hint: you can use the `apply` function across all rows of the dataframe, applying the `sd` function to get a metric of variation.
3. Which species spends the most total time at the tips of branches?
4. Which species pair is most similar in resource use? Is this consistent with Macarthur's published conclusions?

Optional questions for graduate students:

- What is the null expectation for the correlation matrix between species? How would you test whether species are more similar than by chance? Propose a method and try it out. Hint: see the `EcoSimR` package - or consider a row/column swapping randomization of the time matrix, then report observed values relative to quantiles of the null distribution to get p-values.

Ecological niche modeling and Grinnellian niches

Next we will examine how species have different climate requirements, as seen through their Grinnellian niches. To do so we will estimate realized niches from species' geographic occurrences and from climate data obtained from the WorldClim database (`http://worldclim.org/bioclim`). This database provides global maps of average climate for the late 20th century and includes variables such as mean annual temperature and annual precipitation. The species occcurrence data come from the BIEN database (`http://www.biendata.org`) and are collated from dozens of herbaria / surveys.

We first need to make sure we have the `raster` and `maps` package installed to work with the climate maps.

```
library(raster) # if this produces an error, install these packages in Rstudio
```

```
## Loading required package: sp
```

```
library(ggplot2)
library(rasterVis)
```

```
## Loading required package: lattice
```

```
## Loading required package: latticeExtra
```

```
## Loading required package: RColorBrewer
```

```
##
## Attaching package: 'latticeExtra'
```

```
## The following object is masked from 'package:ggplot2':
##
##     layer
```

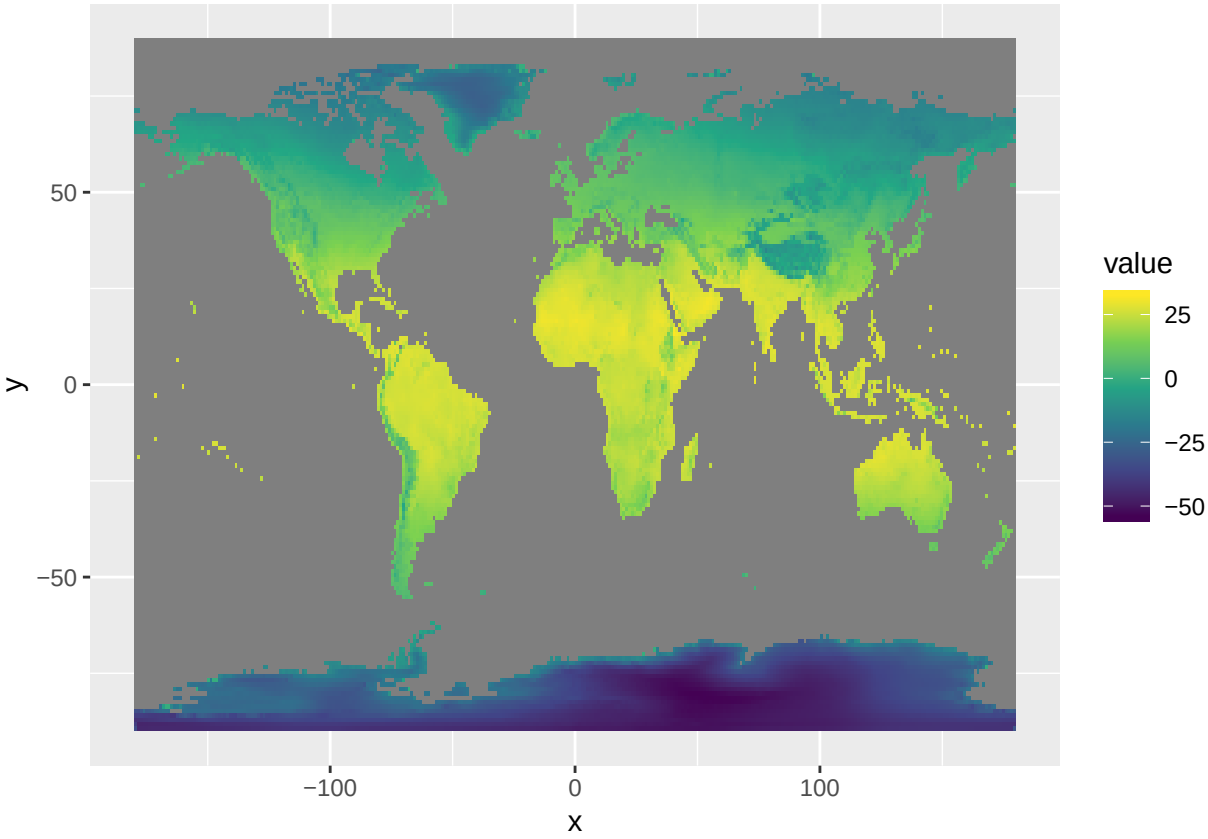
```
# the below data is the BIO_01 variable (mean annual temp, in °C) 10 at 10 arc-minute resolution
# 10 arc-minutes is equal to 0.16 degrees, or using the formula arc length = radius * angle in radians,
# we get 0.16 * pi / 180 * 6371 km) = 18 km at the equator.
raster_mat = raster('wc2.0_bio_10m_01.tif')
raster_mat
```

```
## class       : RasterLayer
## dimensions  : 1080, 2160, 2332800  (nrow, ncol, ncell)
## resolution  : 0.1666667, 0.1666667  (x, y)
## extent      : -180, 180, -90, 90  (xmin, xmax, ymin, ymax)
## coord. ref. : +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0
## data source : /Users/benjaminblonder/Documents/ASU/teaching/bio 423 - spring 2019/BIO 423 labs/Lab 4
## names       : wc2.0_bio_10m_01
## values      : -53.70207, 33.26064  (min, max)
```

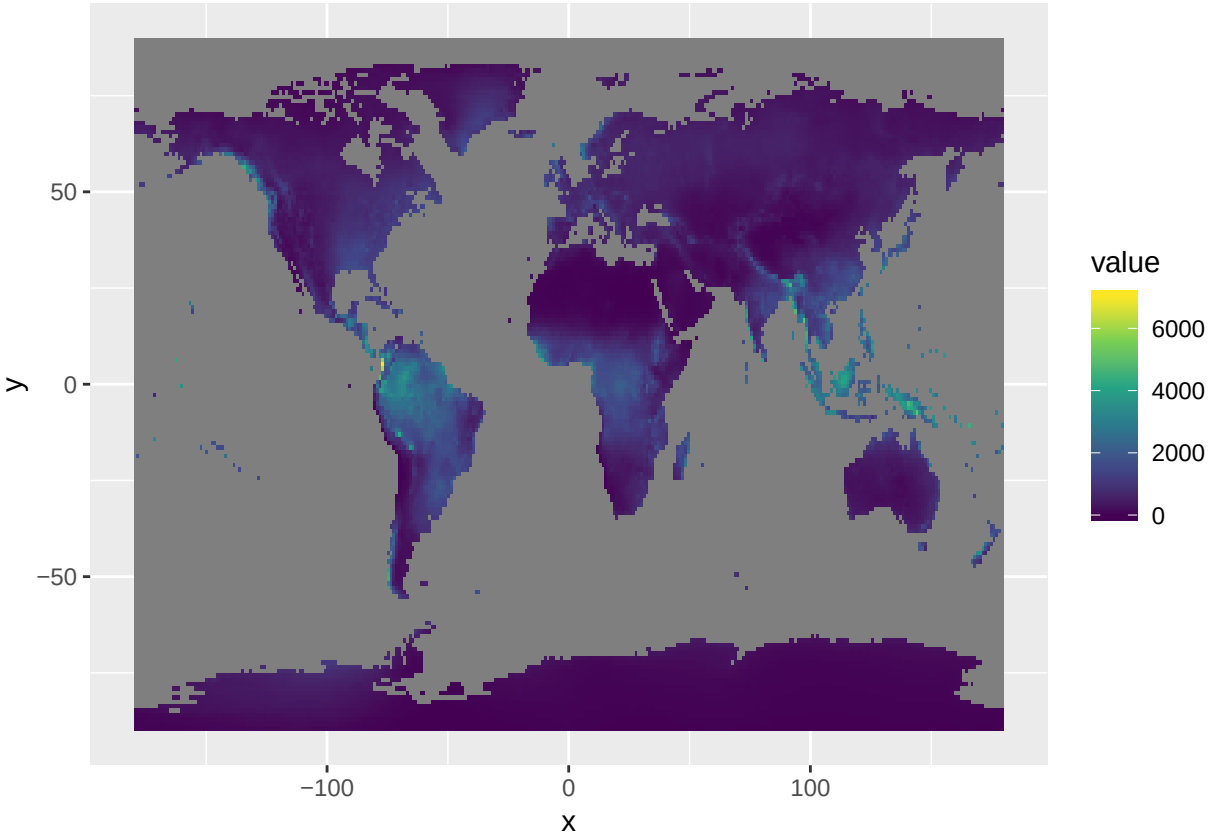
```
# the below data is the BIO_12 variable (annual precipitation, in mm) at the same resolution.
raster_ap = raster('wc2.0_bio_10m_12.tif')
raster_ap
```

```
## class       : RasterLayer
## dimensions  : 1080, 2160, 2332800  (nrow, ncol, ncell)
## resolution  : 0.1666667, 0.1666667  (x, y)
## extent      : -180, 180, -90, 90  (xmin, xmax, ymin, ymax)
## coord. ref. : +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0
## data source : /Users/benjaminblonder/Documents/ASU/teaching/bio 423 - spring 2019/BIO 423 labs/Lab 4
## names       : wc2.0_bio_10m_12
## values      : 0, 11191  (min, max)
```

```
gplot(raster_mat) +
  geom_tile(aes(fill=value)) + scale_fill_continuous(type='viridis')
```




```
gplot(raster_ap) +
  geom_tile(aes(fill=value)) + scale_fill_continuous(type='viridis')
```



Next, we can load in our information on species occurrences. We will specifically be comparing two closely-related oak species, Quercus rubra and Quercus alba.

```
data_oaks = read.csv("oak_distribution.csv")
names(data_oaks) = c("Species","Latitude","Longitude")
str(data_oaks)
```

```
## 'data.frame':    3779 obs. of  3 variables:
##  $ Species  : Factor w/ 2 levels "Quercus alba",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Latitude : num  44.9 45.6 42 44.9 45.5 ...
##  $ Longitude: num  -79.9 -77.4 -82.5 -77.2 -78 ...
```

```
summary(data_oaks)
```

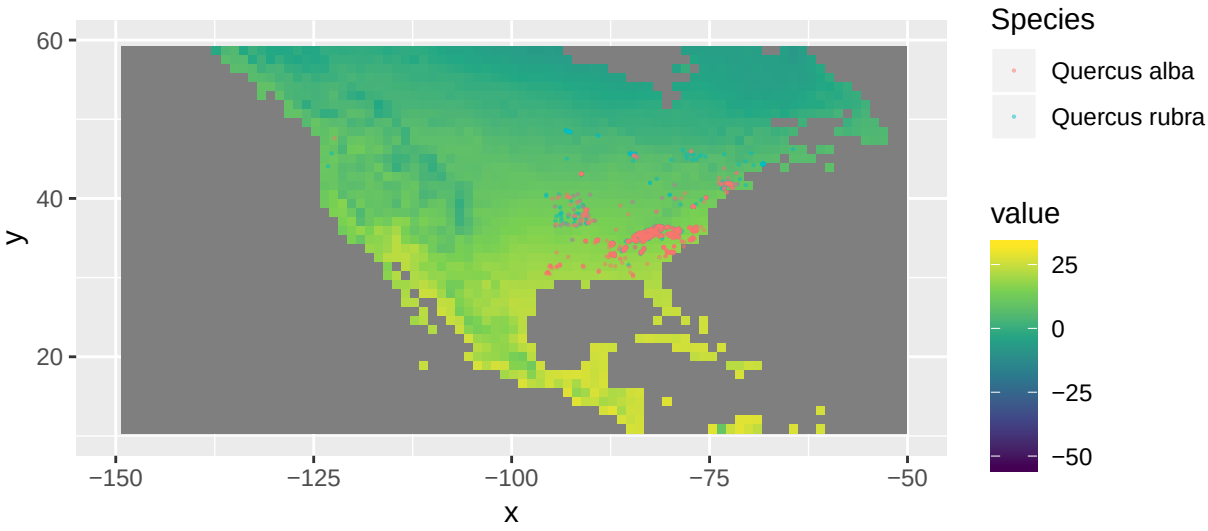
```
##           Species        Latitude       Longitude
##  Quercus alba :1669   Min.   :30.14   Min.   :-123.18
##  Quercus rubra:2110   1st Qu.:35.03   1st Qu.: -83.77
##                       Median :35.37   Median : -83.11
##                       Mean   :35.98   Mean   : -82.90
##                       3rd Qu.:35.85   3rd Qu.: -81.94
##                       Max.   :48.59   Max.   : -64.45
```

We can now combine the occurrence and raster climate data to extract information on the climate at each occurrence of each species. Let's visualize what we are doing, this time using the base R graphics for simplicity:

```
# note we use gplot and not ggplot for rasters... a small quirk of the software
gplot(raster_mat) +
  geom_tile(aes(fill=value)) + scale_fill_continuous(type='viridis') +
  geom_point(data = data_oaks,aes(x=Longitude,y=Latitude,col=Species),alpha=0.5,size=0.1) +
```

```
  xlim(-150,-50) + ylim(10,60) + coord_equal()
```

```
## Warning: Removed 46056 rows containing missing values (geom_tile).
```




```
# extract values from the raster
# to extract values we need to pass in a set of x-y coordinates - here, longitude, and latitude
data_oaks$Mean.Annual.Temperature = extract(raster_mat, data_oaks[,c("Longitude","Latitude")])
data_oaks$Annual.Precipitation = extract(raster_ap, data_oaks[,c("Longitude","Latitude")])

# now each occurrence of each species is associated with a temperature and precipitation value
str(data_oaks)
```

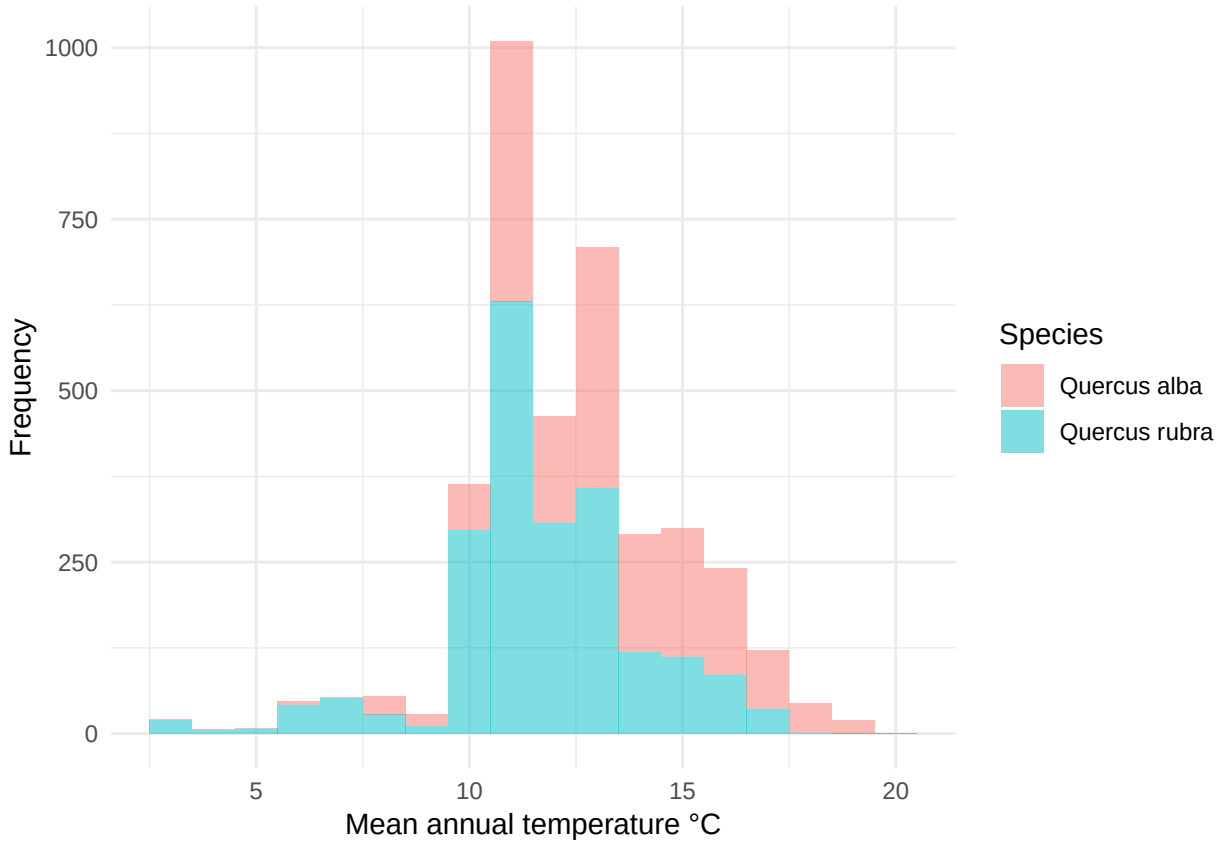
```
## 'data.frame':    3779 obs. of  5 variables:
##  $ Species                : Factor w/ 2 levels "Quercus alba",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Latitude               : num  44.9 45.6 42 44.9 45.5 ...
##  $ Longitude              : num  -79.9 -77.4 -82.5 -77.2 -78 ...
##  $ Mean.Annual.Temperature: num  6.18 4.69 9.76 4.91 3.9 ...
##  $ Annual.Precipitation   : num  1030 851 906 956 872 ...
```

We can summarize the realized niche of each species using these data:

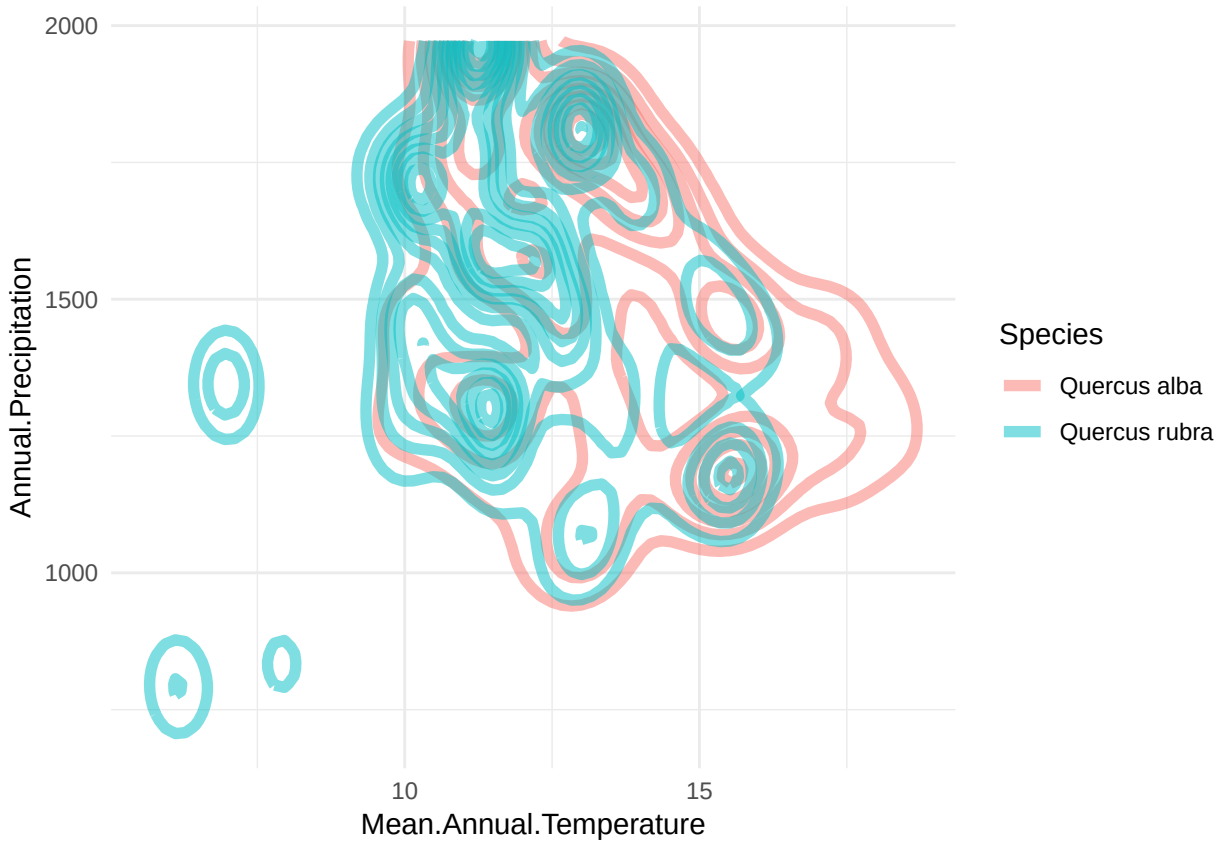
```
# using the apply functions
tapply(data_oaks$Mean.Annual.Temperature, data_oaks$Species, mean)
```

```
##  Quercus alba Quercus rubra
##      13.22701      11.71553
```

```
# visualized with a histogram with 1°C bins
ggplot(data_oaks,aes(x=Mean.Annual.Temperature,fill=Species)) +
  geom_histogram(alpha=0.5,binwidth=1) +
  theme_minimal() +
  xlab("Mean annual temperature °C") + ylab("Frequency")
```




```
# visualized as two-dimensional density plot
ggplot(data_oaks,aes(x=Mean.Annual.Temperature,y=Annual.Precipitation,col=Species)) +
  geom_density_2d(alpha=0.5,size=2) +
  theme_minimal()
```




```
# we can also test for signifcant differences in mean temperatures
# here the p-value is relative to the null hypotheses of the
# two species sharing the same mean temperature preference
t.test(Mean.Annual.Temperature~Species,data=data_oaks)
```

```
##
##  Welch Two Sample t-test
##
## data:  Mean.Annual.Temperature by Species
## t = 20.108, df = 3563.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.364105 1.658854
## sample estimates:
##  mean in group Quercus alba mean in group Quercus rubra
##                    13.22701                    11.71553
```

Questions

5. Which oak species requires a higher annual precipitation? Is this finding consistent with the known ecology of the species (use Google)?
6. Is the difference in mean precipitation among the oak species significant?
7. Which oak species has a broader temperature niche? (Hint: use `tapply` with `sd`)
8. If the United States becomes warmer and drier, which species do you expect will expand its range, and which will contract its range?

Optional questions for graduate students

1. We used a t-test for simplicity. Check if the data meet the assumptions of the t-test, and if not, use another test (e.g. Wilcoxon rank-sum test).
2. You can also build an environmental niche model to predict the potential geographic range of each species. Try installing the `sdm` package and following their tutorial. You will need to choose psuedo-absences for the model, choose an algorithm (try a GLM), and a thresholding method (try a kappa threshold). Which species has a larger potential geographic range?
3. The occurrence data have uneven sampling coverage (note the Missouri-shaped set of occurrence points in the plots). How might this bias your realized niche estimates? What could be done to minimize this bias? (Hint: there is a very big literature on this if you want to learn more!)

What to hand in

- A single Word Document including:
  - written answers (1-2 sentences) and figures for each question above
  - A copy of your R script (the contents of your `.R` file pasted into the Word document)
  - Author contribution statement