Introduction to species distribution modeling for infectious disease vectors*

John M. Drake & Pejman Rohani

Introduction

Arthropod disease vectors are often highly sensitive to environmental conditions such as temperature and precipitation. Thus, transmission of arboviruses and other vector-borne diseases can be very heterogeneous. Many research and policy questions are require modeling the potential distributions of disease vectors. In this exercise, students are introduced to the basic operations of species distribution modeling in R. Upon completion of this exercise, students should be able to

- Read and analyze spatial point data
- Produce choropleth maps
- Fit simple specices distribution models
- Evaluate fit models over different spatial extents

Study system

Several emerging viruses – particularly Chikungunya virus, dengue virus, and Zika virus – are transmitted by the abundant mosquito species *Aedes aegypti* and *Aedes albopictus*. *Ae. aegypti* was historically widespread in the Eastern U.S., but is not primarily concentrated in the South and Gulf Coast. *Ae. albopictus* is an invasive speices and widespread throughout the continent U.S. Despite widespread records, data on both species are still patchily distributed. Thus, it is useful to have a model of the species potential distribution.

Data

County-level mosquito occurrences reported between 1995 and 2016 [1] were combined with point occurrences, collected between 1960 and 2014 [2], aggregated to county-level presence data. Counties with single and multiple occurrences were treated identical. Our final occurrence dataset had 257 and 1478 counties with reported presences for *Ae. aeygpti* and *Ae. albopictus*, respectively. These county-level occurrence records are provides in the file `aedes-data.csv`. This data file also contains the FIPS code,

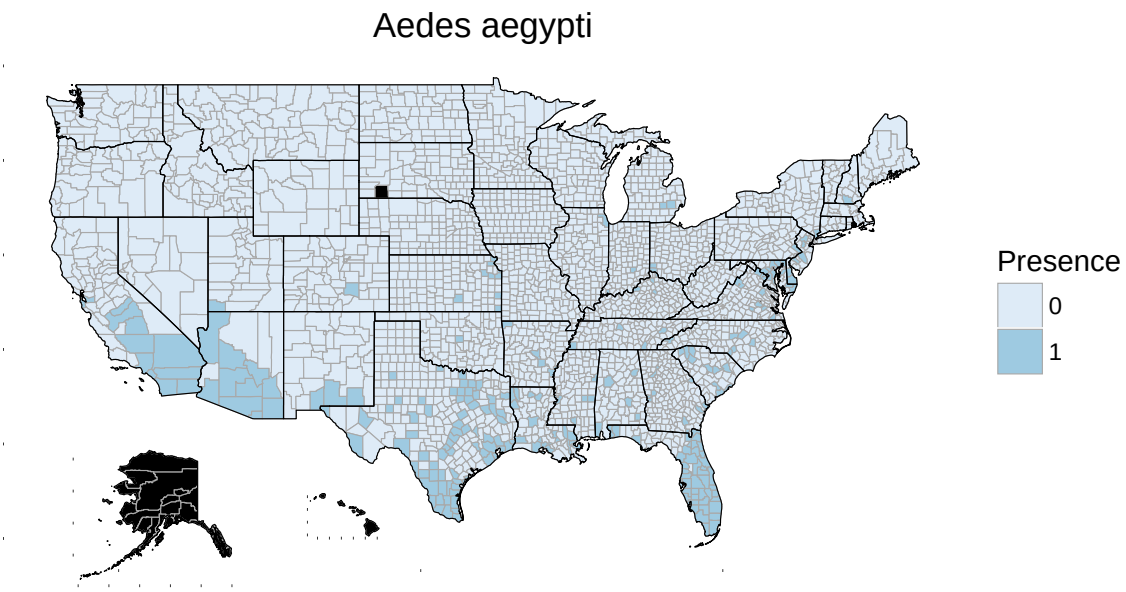
which is a standard key for county identification in the United States, and the geographic coordinates of the county centroid.

In R, the data can be read in using the function `read.csv`.

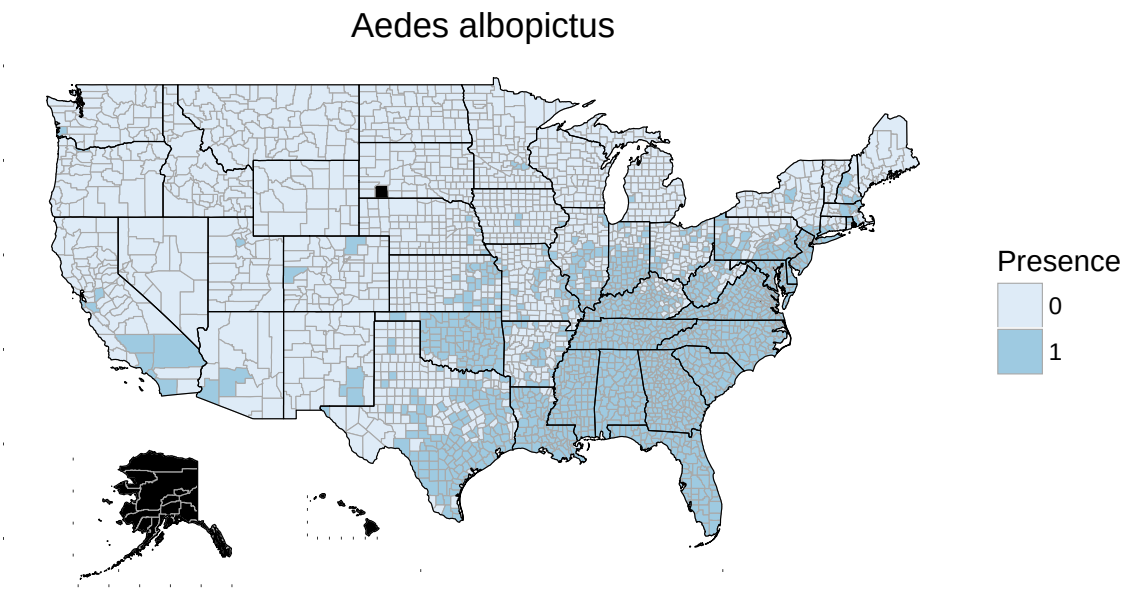
```
> data <- read.csv('aedes-data.csv')
```

First we inspect which counties have records of *Ae. aegypti* and *Ae. albopictus*, respectively. This can be done simply in R using the `choroplethr` package. We also load the package `ggplot2` to facilitate visualization.

```
> library(choroplethr)
> library(ggplot2)
> par(mfrow=c(2,1))
> county_choropleth(data.frame(region=data$FIPS, value=data$Aegypti),
+                   title= "Aedes aegypti",
+                   legend= "Presence", num_colors = 2)
```

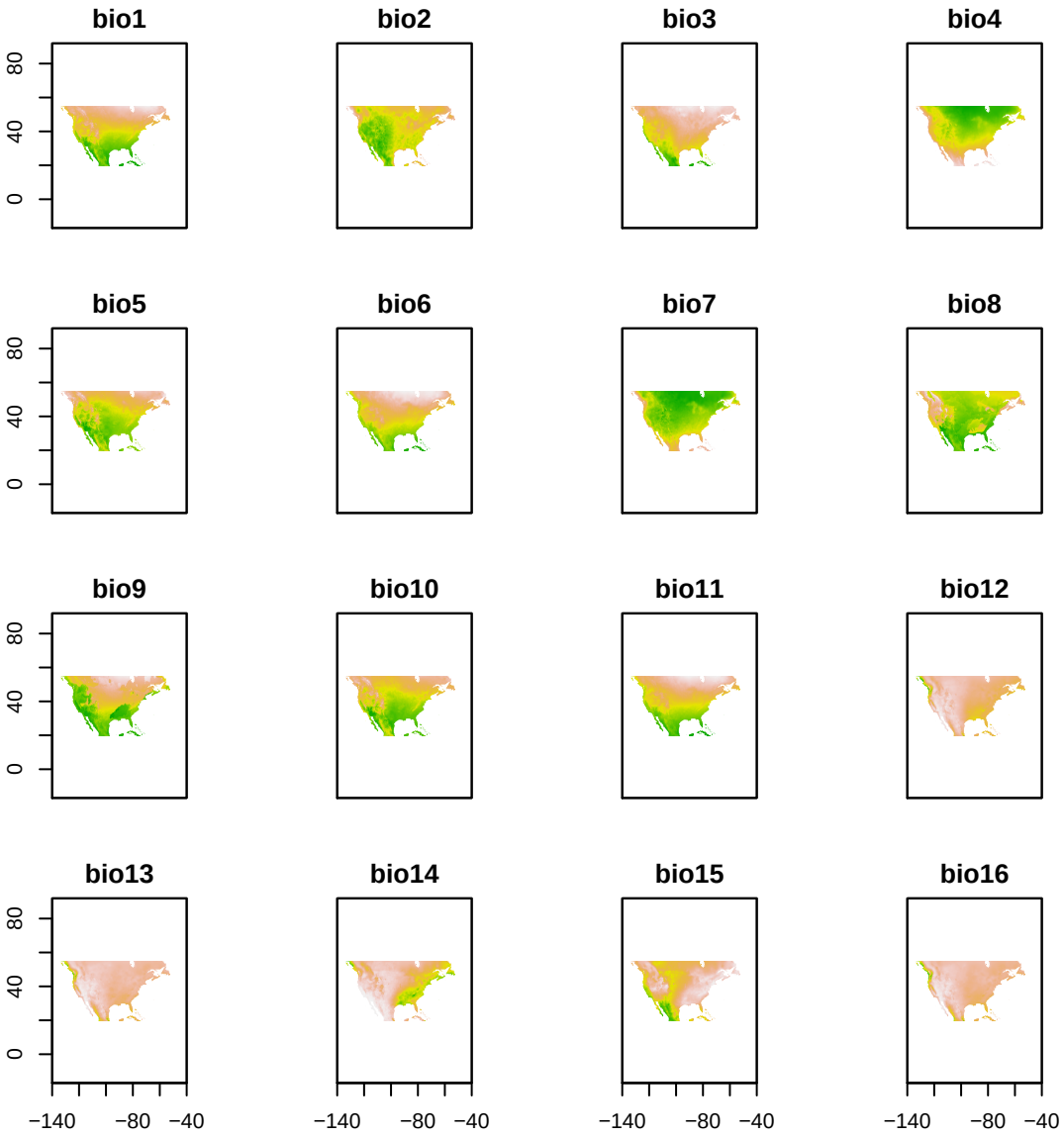



```
> county_choropleth(data.frame(region=data$FIPS, value=data$Albo),
+                   title= "Aedes albopictus",
+                   legend= "Presence", num_colors = 2)
```



These maps are drawn from *shape files*. We will often want to work with grids of environmental data, or *rasters*. Particularly, bioclimatic variables and other environmental layers from the WorldClim project are available in raster form using the `getData` function in the `raster` package. The resulting object is a `RasterStack` at the global extent. Before working with the data, we crop it roughly to the region of the continental United States.

```
> library(raster)
> worldclim <- getData('worldclim', var='bio', res=10)
> newext <- c(-140, -40, 20, 55)
> worldclim <- crop(worldclim, newext)
> plot(worldclim, legend=FALSE)
```

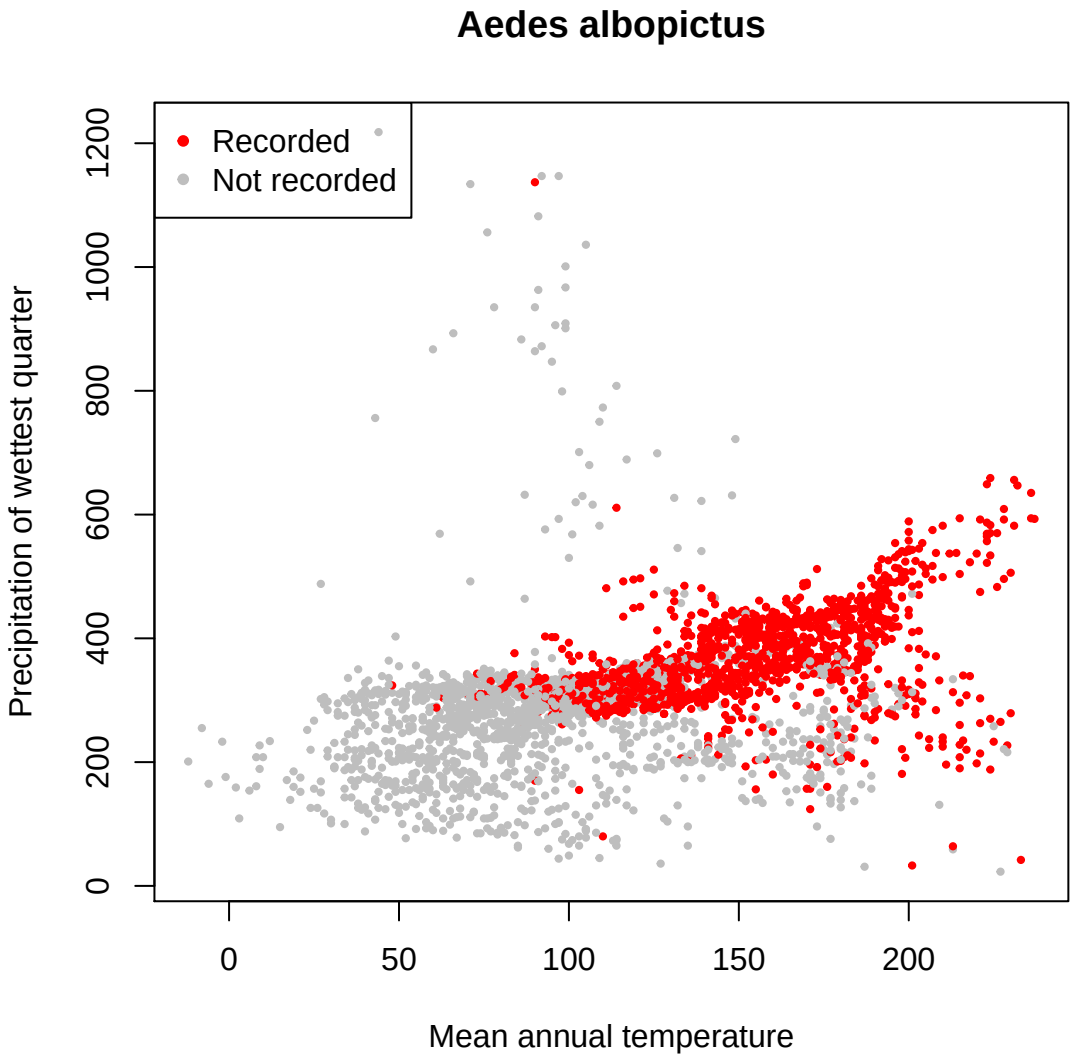


Now we sample the environments in `worldclim` at the county centroids represented in the data. First we create a data frame `z` containing the coordinates of interest.

```
> z <- data[,9:10]
> x <- extract(worldclim, z)
```

For illustration, we look at `bio1` (annual mean temperature) and `bio16` (precipitation of the wettest quarter). Red points are environments where *Aedes albopictus* occurs.

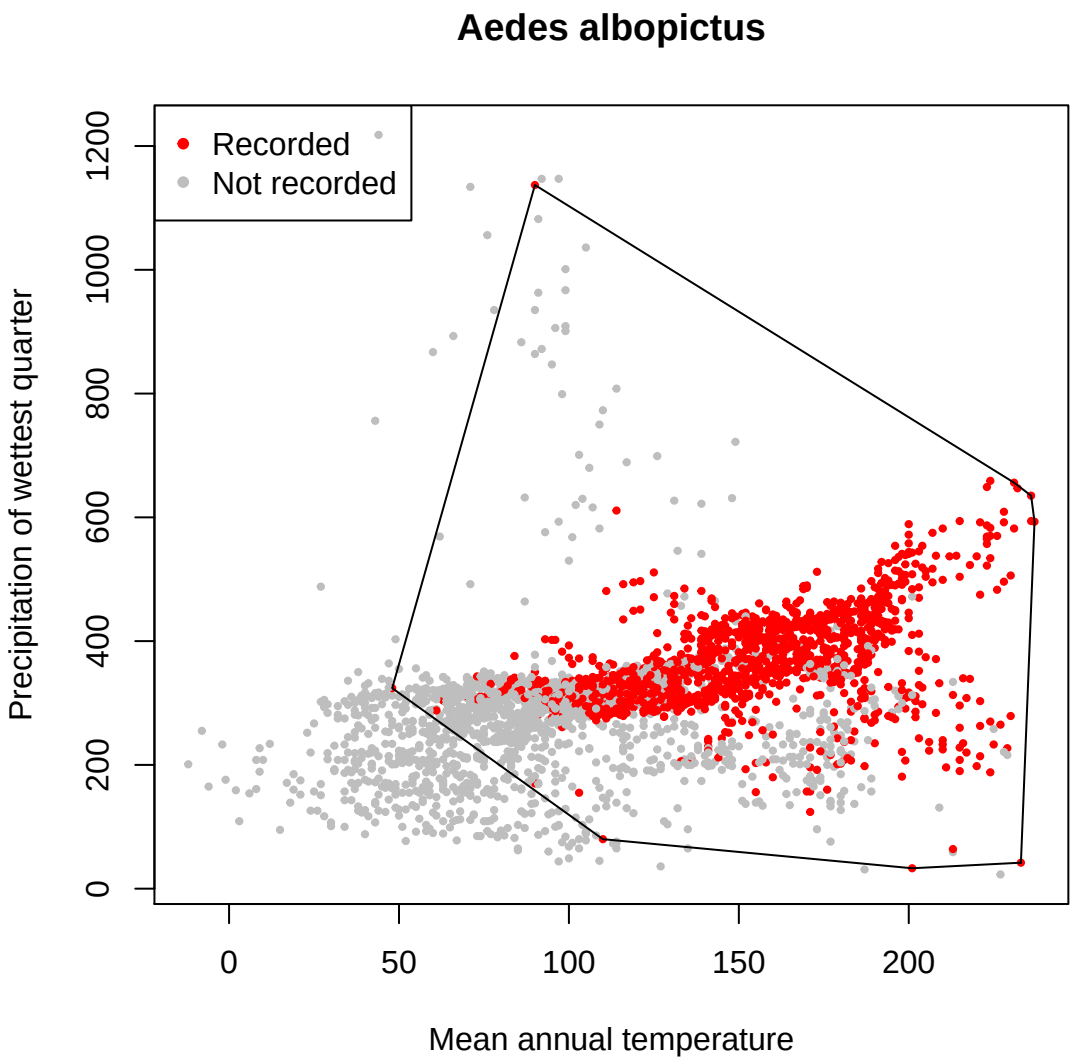
```
> plot(x[,1], x[,16], cex=0.65, pch=20,
+      col=ifelse(data$Albo==1, 'red', 'grey'),
+      xlab='Mean annual temperature',
+      ylab='Precipitation of wettest quarter',
+      main='Aedes albopictus')
>      legend('topleft', col=c('red','grey'), legend=c('Recorded', 'Not recorded'), pch=20)
```



At this point, the number of paths forward increases dramatically. How the project progresses may be guided by modeler preference and intuition, knowledge of different techniques, information in the data, and model goals. One simple approach to *ecological niche modeling* is to find the convex hull containing all occurrence points. The convex hull is the smallest area polygon with no interior angles containing all points. In R, we fit the convex hull using the function `chull`. We add the convex hull, which may be interpreted as a model of the ecological niche.

```
> pts <- data.frame(bio1=x[data$Albo==1,1], bio16=x[data$Albo==1,16])
> model1 <- chull(pts) # this finds the points on the hull
> model1 <- c(model1, model1[1])  # this "closes" the hull
> plot(x[,1], x[,16], cex=0.65, pch=20,
+      col=ifelse(data$Albo==1, 'red', 'grey'),
+      xlab='Mean annual temperature',
+      ylab='Precipitation of wettest quarter',
+      main='Aedes albopictus')
```

```
>      legend('topleft', col=c('red','grey'), legend=c('Recorded', 'Not recorded'), pch=20)
> lines(pts[model1, ])
```



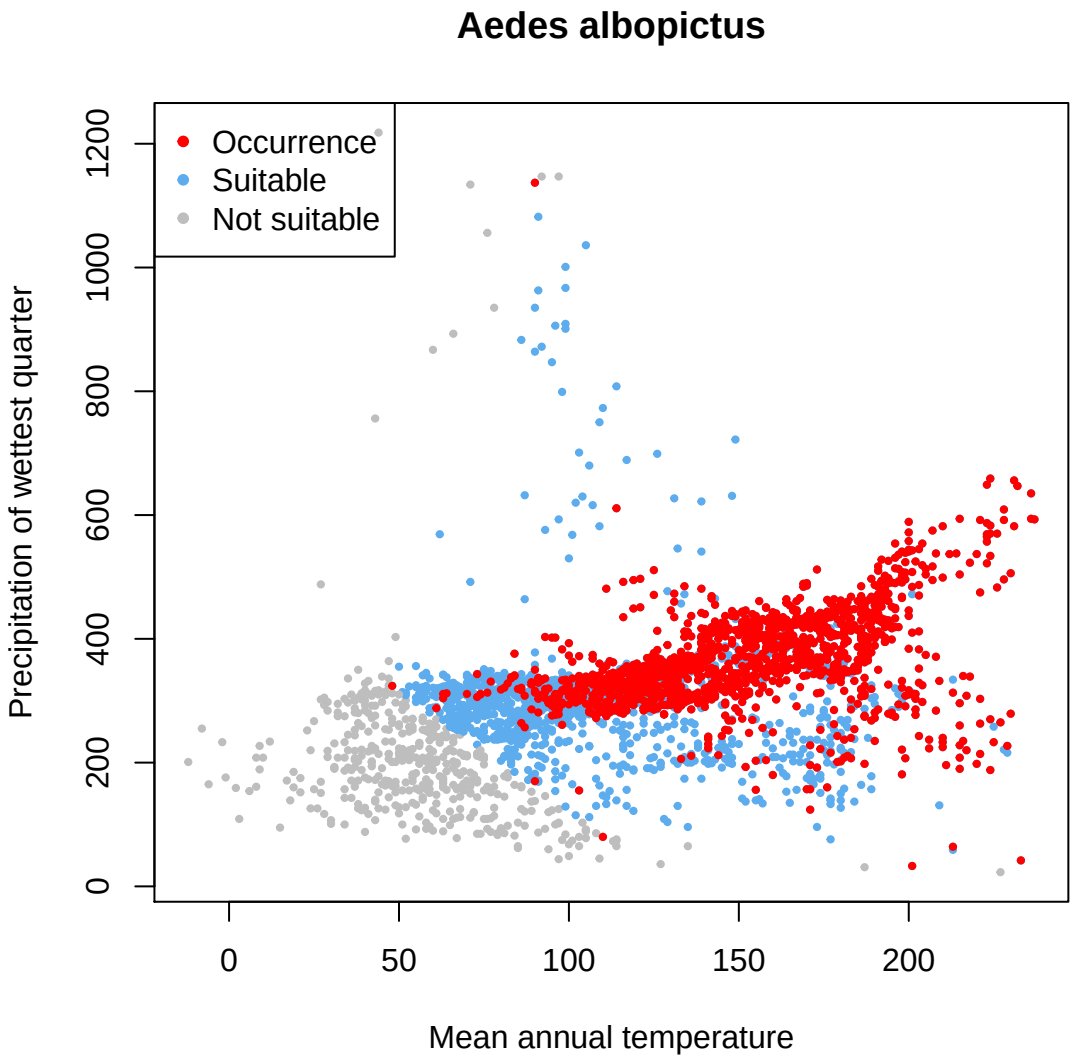
Evidently, there are a lot of *possible environments* that are not realized within this space, but are predicted by this model to be suitable.

Exercise 1. Plot the *Ae. albopictus* niche with respect to two different environmental variables.

Exercise 2. Plot the *Ae. aegypti* niche.

Now, we proceed to ask about the potential habitats of *Ae. albopictus*. For all points, we can test whether or not they fall within the convex hull using the function `tsearchn` from the `geometry` package. (Actually testing whether or not a point falls within a convex hull is not especially straightforward.) The following plot shows the centroid environment at locations where *Ae. albopictus* was recorded, locations where it was not recorded but are predicted to be habitable, and locations outside the niche.

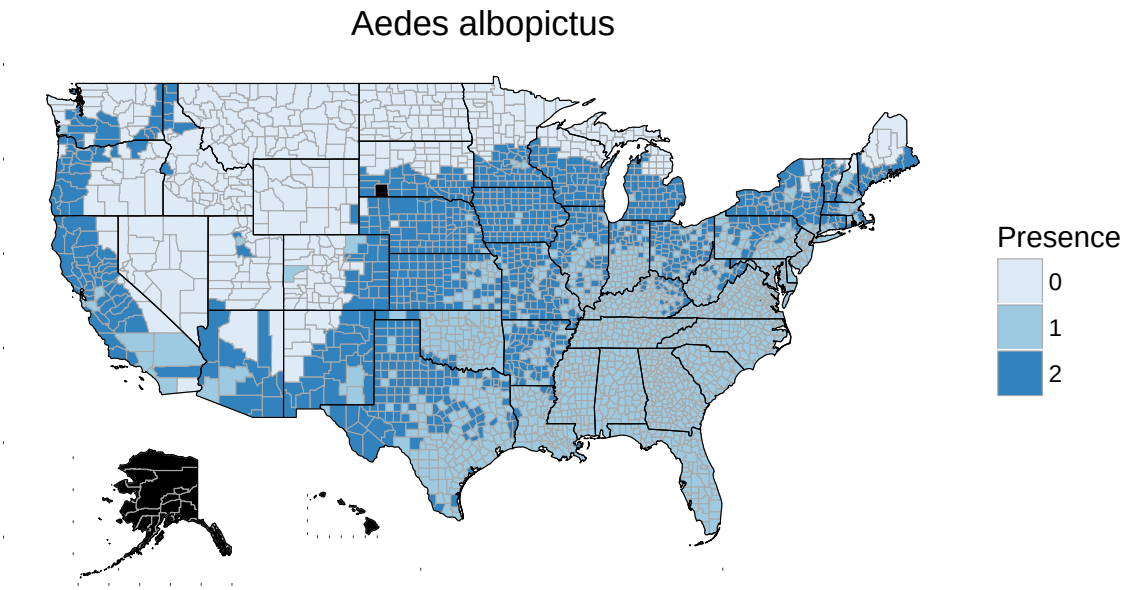
```
> library(geometry)
> pts.test <- data.frame(bio1=x[,1], bio16=x[,16])
> tri.pts <- tsearchn(as.matrix(pts), delaunayn(pts), as.matrix(pts.test))
> test.pts <- !is.na(tri.pts$p[,1])
> plot(x[,1], x[,16], cex=0.65, pch=20,
+      col=ifelse(test.pts, 'steelblue2', 'grey'),
+      xlab='Mean annual temperature',
+      ylab='Precipitation of wettest quarter',
+      main='Aedes albopictus')
> points(x[data$Albo==1,1], x[data$Albo==1,16], col='red',pch=20, cex=0.65)
> legend('topleft', col=c('red','steelblue2','grey'), legend=c('Occurrence', 'Suitable', 'Not suitable'
```



We can also plot these on a county-level map.

```
> data$AlboPredicted <- ifelse(data$Albo==1,1,ifelse(test.pts,2,0))
```

```
> county_choropleth(data.frame(region=data$FIPS, value=as.factor(data$AlboPredicted)),
+                   title= "Aedes albopictus",
+                   legend= "Presence", num_colors = 3)
```



Exercise 3. Plot the *Ae. aegypti* niche on a county-level map.

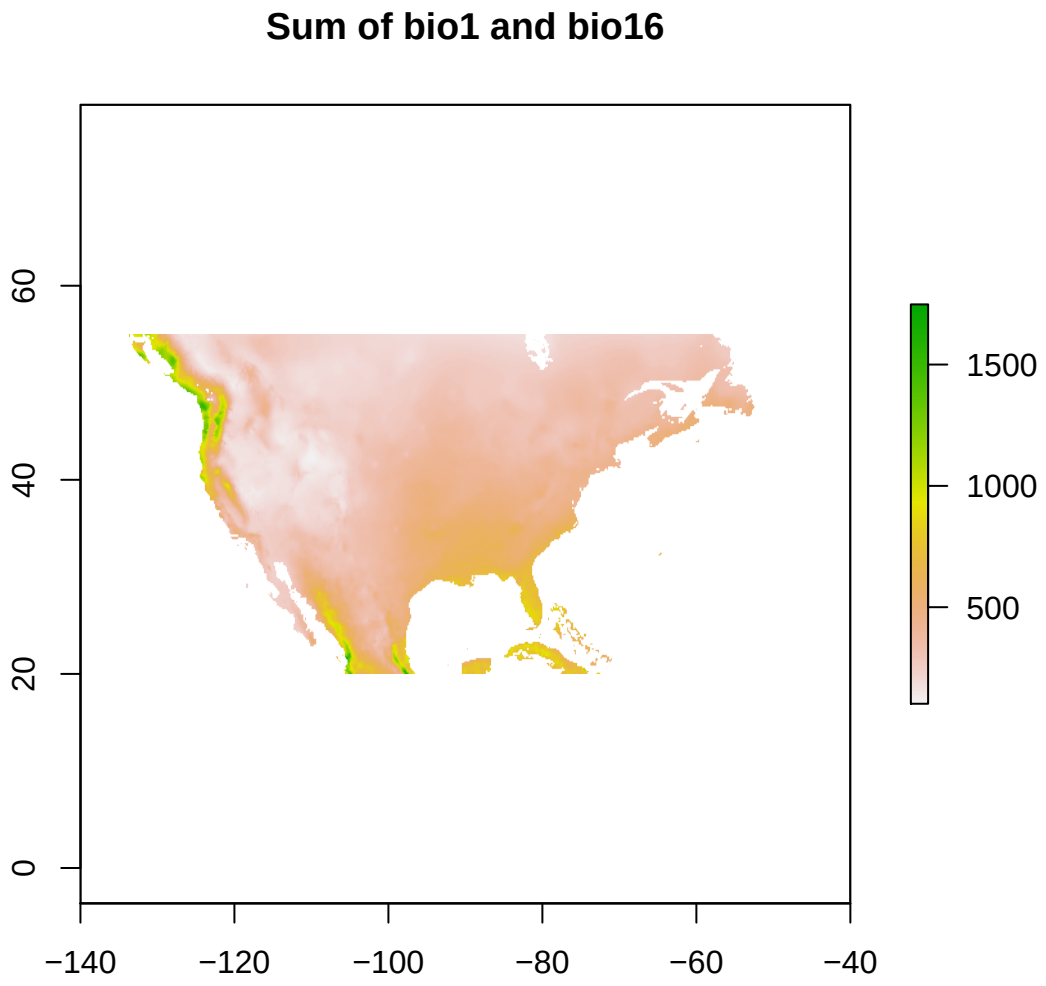
***Exercise 4.** Devise a scheme to pick the combination of two variables that best describes the niche. (Consider, what might we mean by "best"?) Implement this scheme and plot the best niche model.

In the final stage of this exercise, we seek to plot the niche at a finer resolution, using the worldclim data. That is, we wish to test each point in the raster to determine if it belongs to the *Ae. albopictus* niche. First we make a raster stack (called `shortstack`) comprising just `bio1` and `bio16`.

```
> shortstack <- subset(worldclim, c(1,16))
```

In general, we can use the function `calc` to perform a computation over a raster stack. For instance, here we illustrate `calc` by computing and plotting the sum of `bio1` and `bio16`. (Note: this is a meaningless quantity, calculated here just for the sake of illustration.)

```
> raster.sum <- calc(shortstack, fun=sum)
> plot(raster.sum, main='Sum of bio1 and bio16')
```



Here we write a function (called `rasterhull`) that takes two values (i.e. the values for the two layers in `shortstack`) and test whther the point falls within the compuited convex hull.

```
> rasterhull <- function(x){
+   # function to test whether an individual raster point x is in the convex hull of pts
+   pts.test <- data.frame(x[1], x[2])
+   tri.pts <- tsearchn(as.matrix(pts), delaunayn(pts), as.matrix(pts.test))
```

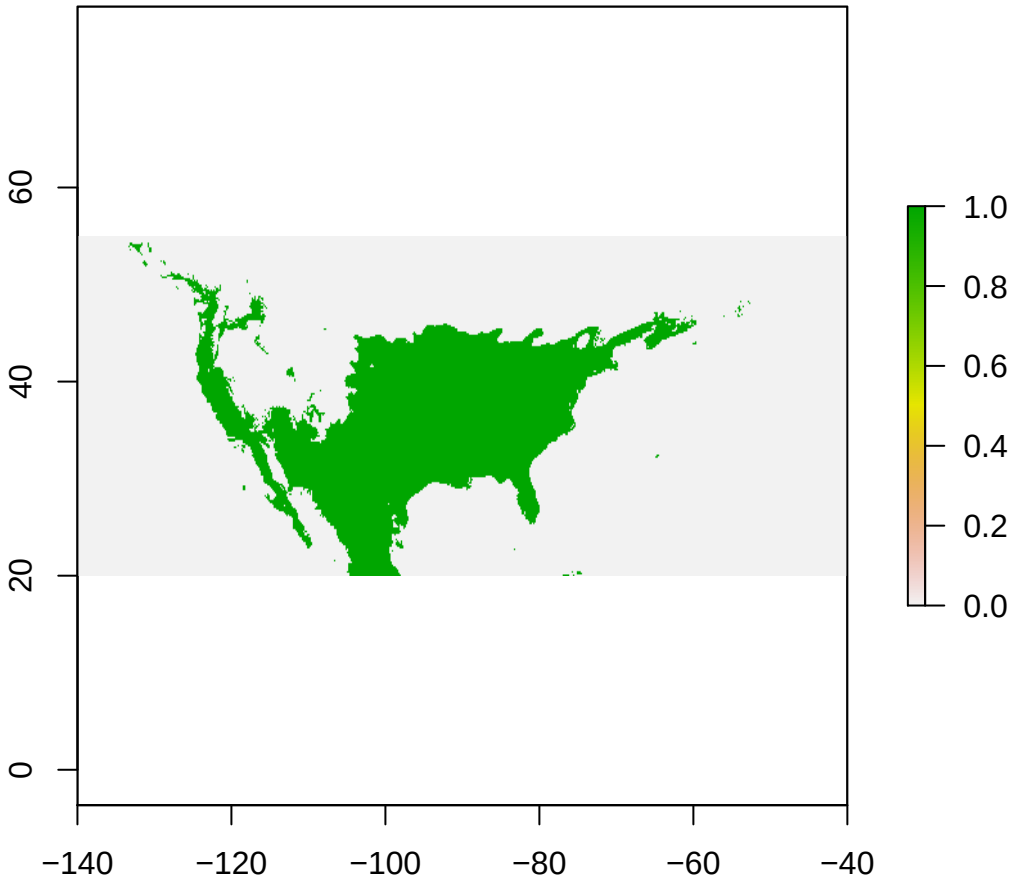
```
+   test.pts <- !is.na(tri.pts$p[,1])
+ }
```

The next step is to call `rasterhull` using `calc`. This takes awhile (15 minutes on my high performance workstation), so I have pre-computed the result and stored it as a raster file called `map-model.grd`.

```
> map.model <- calc(shortstack, fun=rasterhull)
> writeRaster(map.model, filename="map-model.grd", overwrite=TRUE)
```

We can simply load the stored file from disk and plot.

```
> map.model <- raster('map-model.grd') #load map
> plot(map.model)
```



***Exercise 5.** Map at high resolution the potential distribution of *Ae. aegypti*.

References

[1] Hahn MB, Eisen RJ, Eisen L, Boegler KA, Moore CG, McAllister J, et al. Reported Distribution of Aedes (Stegomyia) aegypti and Aedes (Stegomyia) albopictus in the United States, 1995-2016 (Diptera: Culicidae). Journal of Medical Entomology. 2016; p. tjw072. doi:10.1093/jme/tjw072.

[2] Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Brady OJ, et al. The global compendium of Aedes aegypti and Ae. albopictus occurrence. Scientific Data. 2015;2:150035–8.