Lecture 23

Species Distribution Models

GEOG 489

SPRING 2020

Species Distribution Models

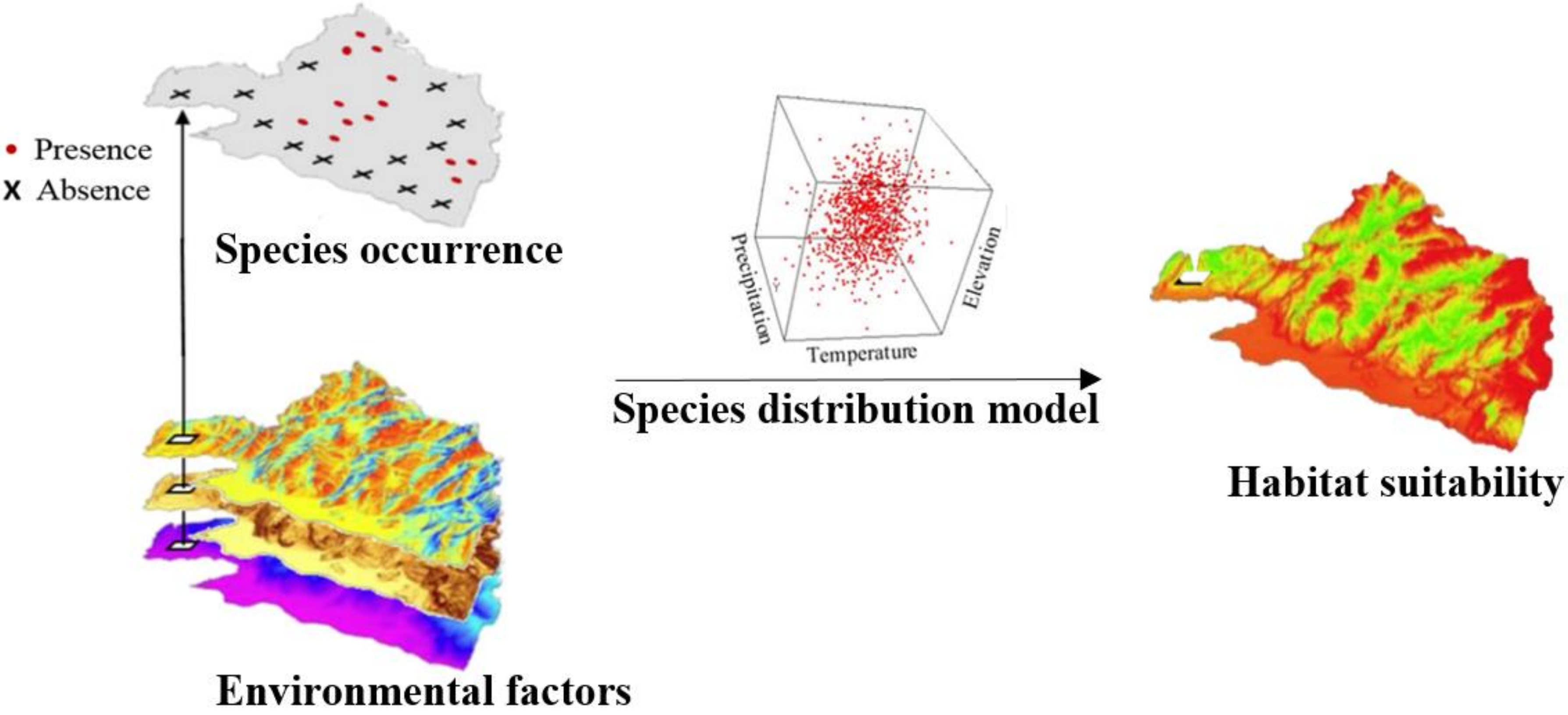
There is a class of models which are developed to predict the probability of something being present at a given location.

These are typically used to predict species presence/absence, so we will refer to them as species distribution models

But note that this class of model can be used for just about any continuous variable.

Species Distribution Models

- Investigate the relationship between species occurrence and environmental gradients to predict the habitat suitability



Species Distribution Models

The basic process of developing an SDM is:

1) Collect data on the presence/absence of a species (typically) with field work at specific locations x, y. Code these as 0 (absent) or 1 (present).

2) Develop raster/vector surfaces that may be related to the distribution of a species.

3) Extract the raster/vector info at each location where the presence/absence data was collected.

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

5) Perform accuracy assessment/goodness of fit analysis.

6) (Optional) tweak the model/inputs and repeat steps 1-5.

7) Apply the model to the raster/vector layers to produce a map of species distributions.

Species Distribution Models

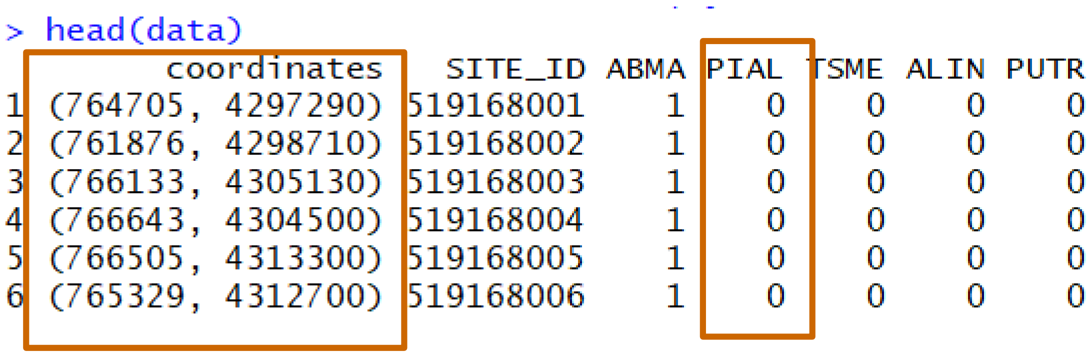
Question: what is the potential distribution of rare plant species Pinus albicaulis (Whitebark pine) that is found in the high Sierra Nevada Mountains?

1) Collect data on the presence/absence of a species (typically) with field work at specific locations x, y. Code these as 0 (absent) or 1 (present).

# 712 observations (if PIAL = 1, a Whitebark pine was present at that location, and if PIAL = 0, the species was absence).

data<-read.csv("tahoe_spp_subset.csv",head=T)

summary(data)



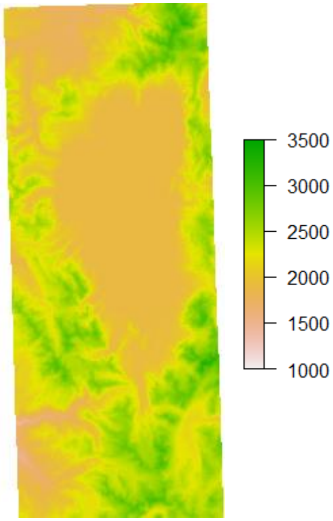
```
> head(data)
        coordinates   SITE_ID ABMA PIAL TSME ALIN PUTR
1 (764705, 4297290) 519168001    1    0    0    0    0
2 (761876, 4298710) 519168002    1    0    0    0    0
3 (766133, 4305130) 519168003    1    0    0    0    0
4 (766643, 4304500) 519168004    1    0    0    0    0
5 (766505, 4313300) 519168005    1    0    0    0    0
6 (765329, 4312700) 519168006    1    0    0    0    0
```

Species Distribution Models

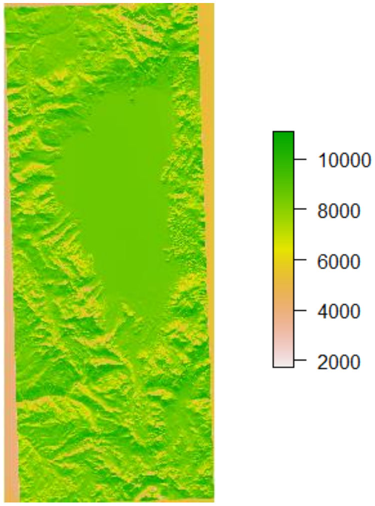
2) Develop raster/vector surfaces that may be related to the distribution of a species.

We have a number of possible environmental/topographic predictors of Whitebark pine, including elevation, radiation, and topographic convergence index (TCI)

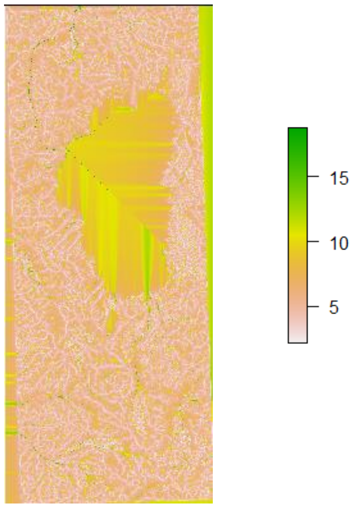
```
elev <- raster("tahoedems_nad83_30m.img")
rad <- raster("tahoe_rad_reproj.tif")
tci <- raster("tahoedems_tci.img")
```



Elevation



Radiation

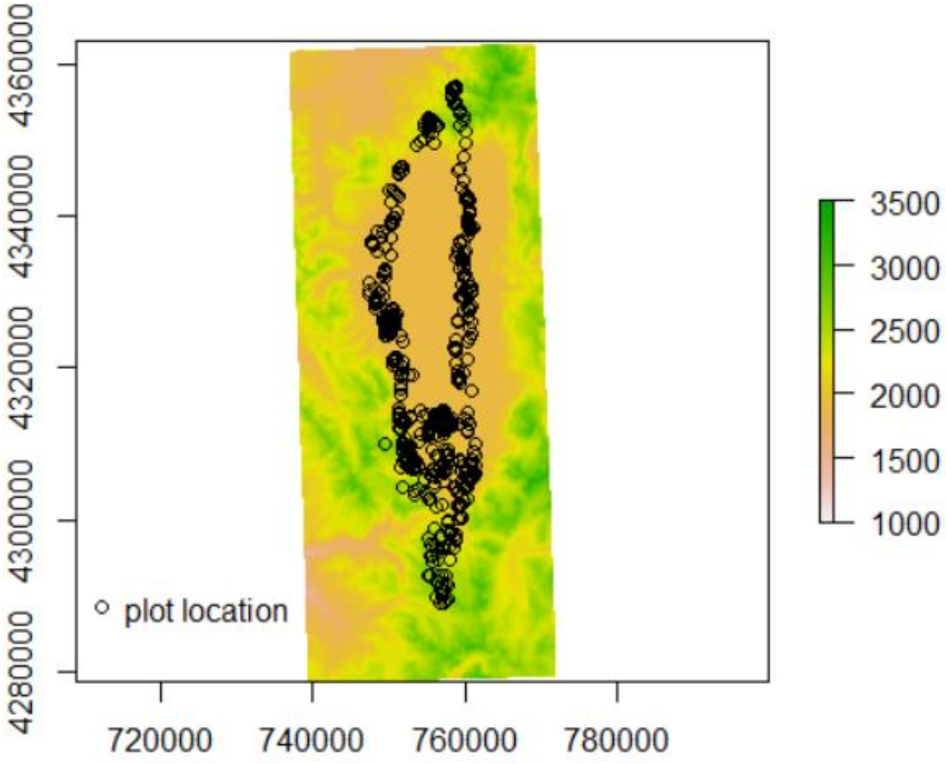


TCI

Species Distribution Models

Species distribution – Environmental predictors

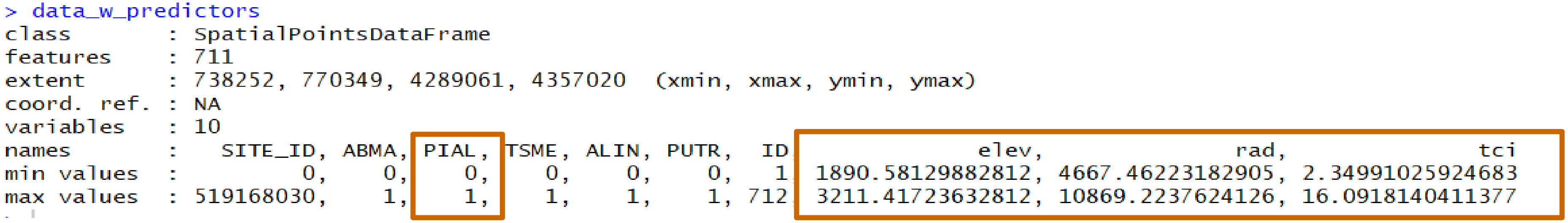
# 712 observations (presence: 58; absence: 654)



Species Distribution Models

3) Extract the raster/vector info at each location where the presence/absence data was collected.

extracted_predictors <- extract(predictor_stack,data,df=TRUE)



```
> data_w_predictors
class       : SpatialPointsDataFrame
features    : 711
extent      : 738252, 770349, 4289061, 4357020  (xmin, xmax, ymin, ymax)
coord. ref. : NA
variables   : 10
names       :   SITE_ID, ABMA, PIAL, TSME, ALIN, PUTR,  ID             elev,              rad,              tci
min values  :         0,    0,    0,    0,    0,    0,   1 1890.58129882812, 4667.46223182905, 2.34991025924683
max values  : 519168030,    1,    1,    1,    1,    1, 712 3211.41723632812, 10869.2237624126, 16.0918140411377
```

Species Occurrence

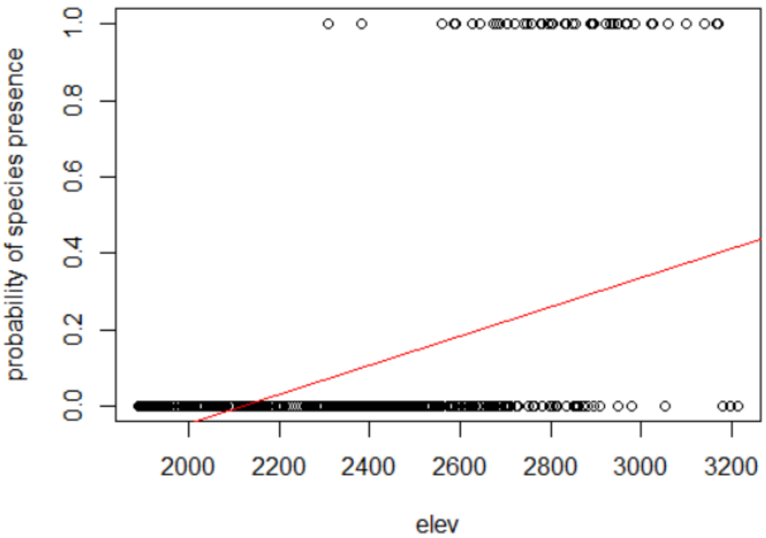
Environmental Predictors

Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

We have a data set where if PIAL = 1, a Whitebark pine was present at that location, and if PIAL = 0, the species was absence. So, what if we just regress this field against, say, elevation:

```
lm_notransform <- lm(PIAL ~ elev,data=data_w_predictors)
```



Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

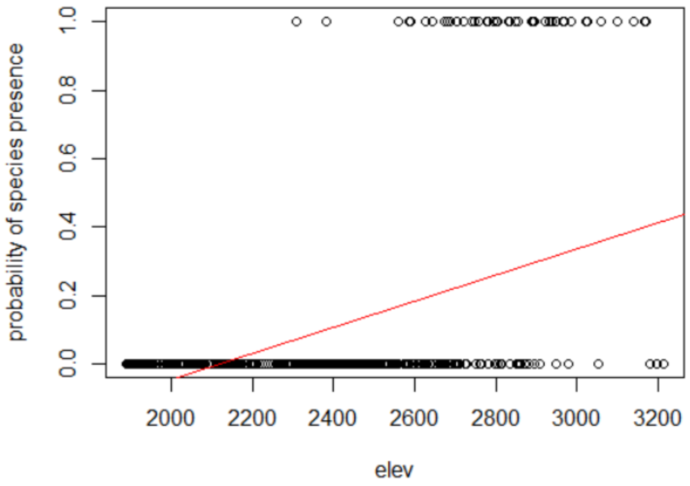
# Some notes about using linear regression:

1) The input data isn't really continuous.

2) The linear model predicts probabilities < 0 or > 1, which is not possible.

3) The variance isn't constant across X.

4) Significance test of the regression coefficients assumes the errors in prediction are normally distributed (which is clearly not the case).



Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

Logistic regression

We can take advantage of a generalized linear models (glm) and define a different error distribution, in this case, a binomial distribution.

To fit this distribution, we use a "logit link function“ to transform our input data such that: ln(P / (1-P)) = a + bX

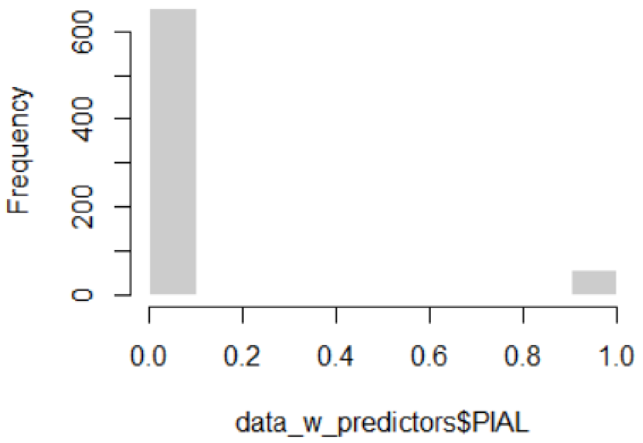
$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

P is the probability of species being present in a location

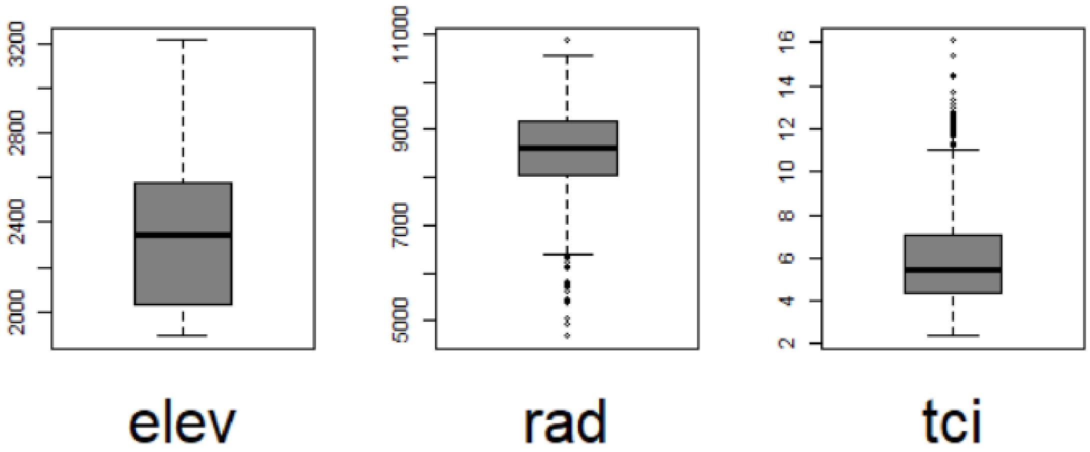
Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

Logistic regression – descriptive stats



Species Occurrence



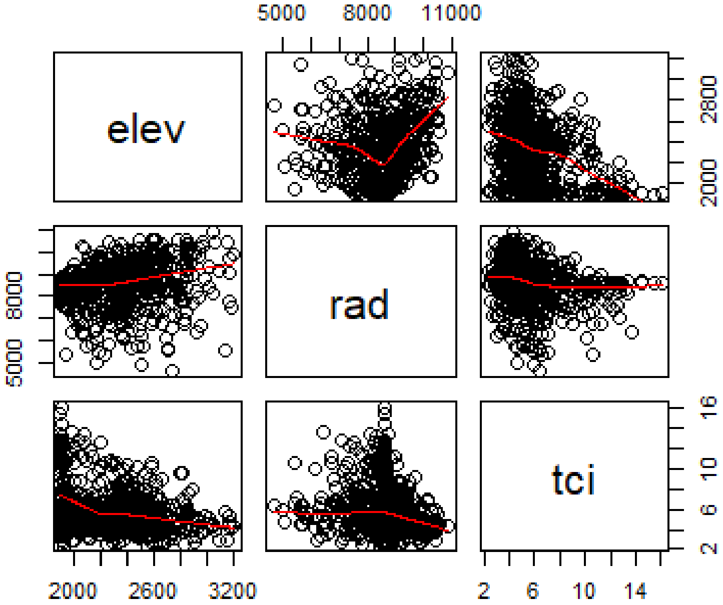
Environmental Predictors

Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

Logistic regression – descriptive stats

# GLMs really need the variables to be (somewhat) uncorrelated. Let's look at the pairwise correlations:

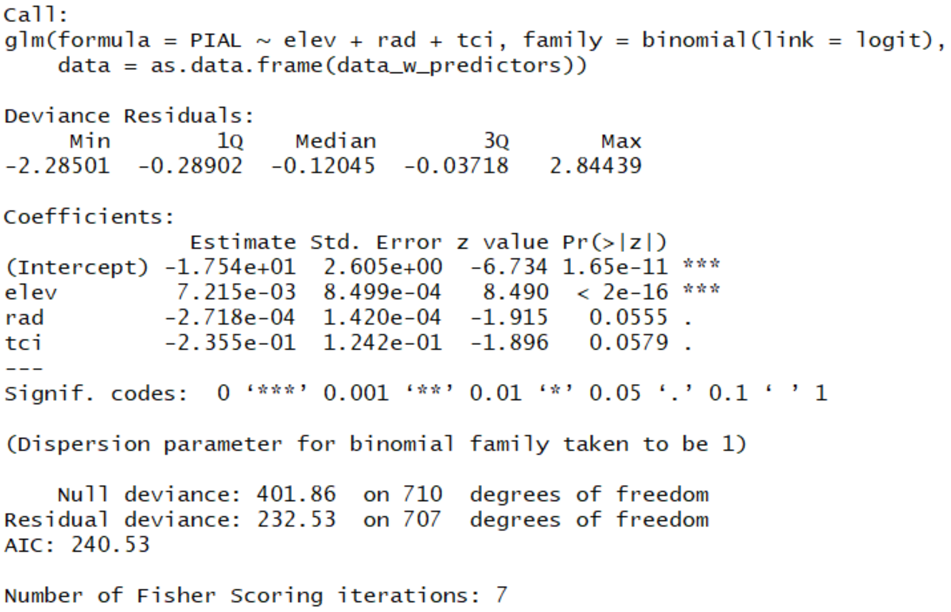


Species Distribution Models

4) Construct a model relating the probability of presence as a function of the extracted raster/vector variables.

Logistic regression: species occurrence ~ environmental predictors

pial.sdm <- glm(PIAL~elev+rad+tci, family=binomial(link=logit), data = as.data.frame(data_w_predictors))



```
Call:
glm(formula = PIAL ~ elev + rad + tci, family = binomial(link = logit),
    data = as.data.frame(data_w_predictors))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28501  -0.28902  -0.12045  -0.03718   2.84439

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.754e+01  2.605e+00  -6.734 1.65e-11 ***
elev         7.215e-03  8.499e-04   8.490  < 2e-16 ***
rad         -2.718e-04  1.420e-04  -1.915   0.0555 .
tci         -2.355e-01  1.242e-01  -1.896   0.0579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 401.86  on 710  degrees of freedom
Residual deviance: 232.53  on 707  degrees of freedom
AIC: 240.53

Number of Fisher Scoring iterations: 7
```

Some things to note:

1) The coefficients allow us to see the impacts of the predictor in the probability of its presence/absence.

2) Elevation has a significant, positive relationship on PIAL being present. This means that as elevation increases, the probability of finding that species increases.

3) TCI and radiation are less significant, and also have small negative coefficients.

Species Distribution Models

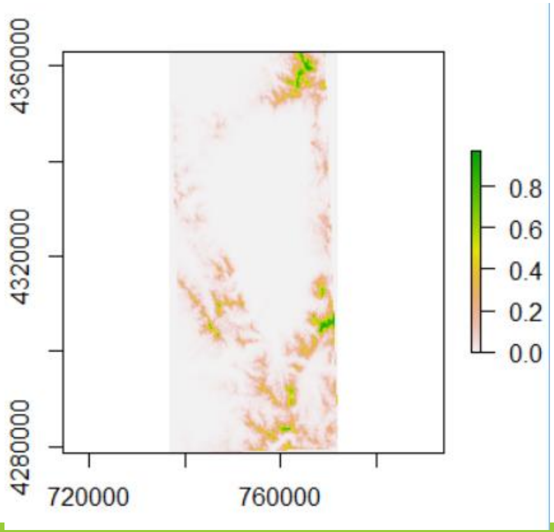
5) Perform accuracy assessment/goodness of fit analysis (not in R file).

6) (Optional) tweak the model/inputs and repeat steps 1-5 (not in R file).

7) Apply the model to the raster/vector layers to produce a map of species distributions.

```
# Now, we can apply this model to our predictor rasters!

pial.surf <- predict(predictor_stack, pial.sdm, type="response")
```



Probability of species occurrence

Species Distribution Models

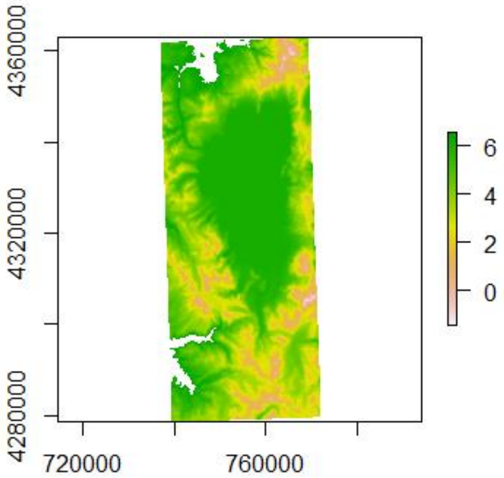
One of the applications of SDMs is to look at climate change impacts on species.

First, let's use an actual climate variable, not a topographic one. A simple model is that temperature decreases with elevation. We can use a lapse rate model to describe this.

At the lake shore, annual average temperatures are around 6 deg C. The lake is at 1900m above sea level. The elevational lapse rate was found to be 5.3 deg C/km

We can convert our elevation map to temperature by:

```
reference_elevation = 1.900
reference_temperature = 6
lapse_rate = 5.3 # From Dobrowski et al. 2009
temperature = reference_temperature -
  (lapse_rate*(elev/1000 - reference_elevation))
temperature[elev < 1800] <- NA
```



Temperature

Species Distribution Models

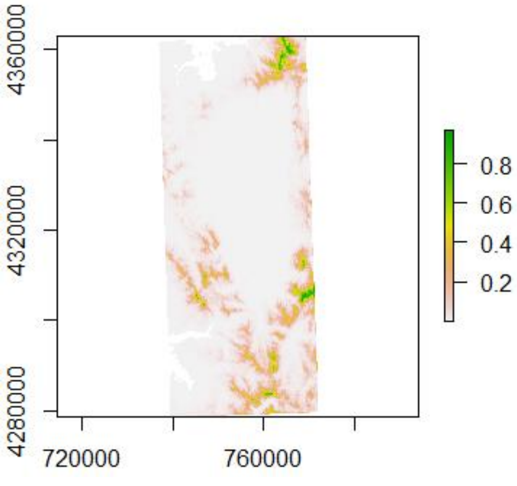
One of the applications of SDMs is to look at climate change impacts on species.

```
# Logistic regression: (species occurrence ~ temperature + radiation + TCI)

pial.sdm <- glm(PIAL~temperature+rad+tci, family=binomial(link=logit),
data=as.data.frame(data_w_predictors))

# Predict species distributions with current temperature

pial.surf_wtemp <- predict(predictor_stack,pial.sdm,type="response")
```



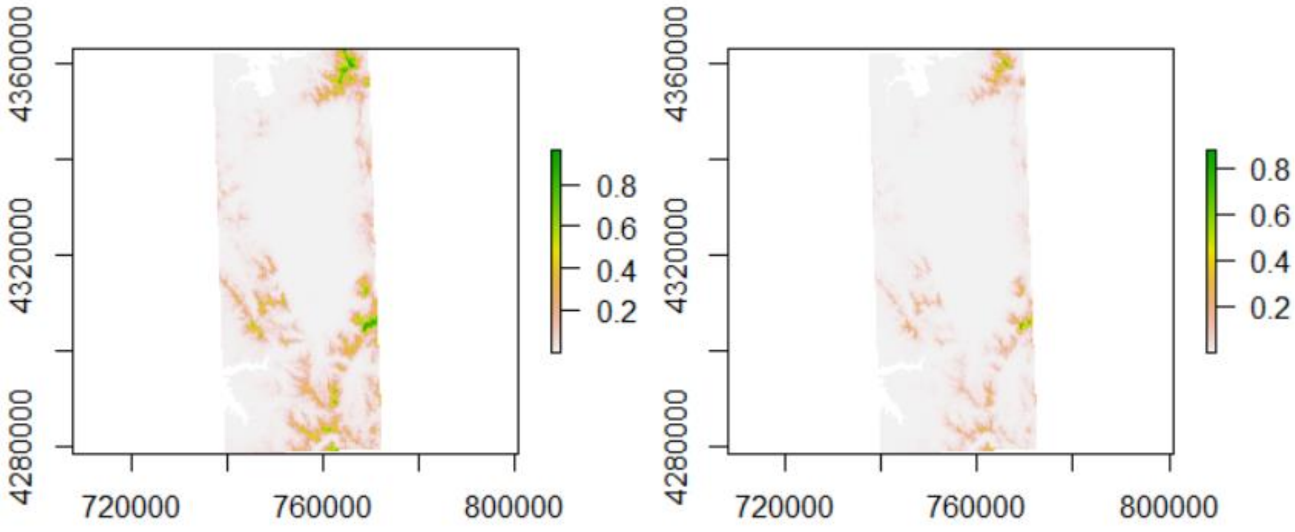
Species distribution with current temperature

Species Distribution Models

One of the applications of SDMs is to look at climate change impacts on species.

```
# Let's simulate a 1.0 deg C regional warming:
temperature_future <- temperature + 1.0
future_stack <- stack(temperature_future,rad,tci)

# Predict species distributions with current temperature
pial.surf_wtemp <- predict(future_stack, pial.sdm, type="response")
```



Species distribution with current temperature (left) and future climate change (right)

Species Distribution Models

One of the applications of SDMs is to look at climate change impacts on species.

```
# Notice the probabilities decreased. Let's calculate the change in area. We'll assume
that a prob > 0.7 is "Whitebark pine habitat":

# Current area  (1053.9 hectares)

present_PIAL <- rasterToPolygons(pial.surf_wtemp, function(x) { x > 0.7 },dissolve=TRUE)

sum(sapply(slot(present_PIAL, "polygons"), slot, "area"))*0.0001

# Future area  (78.57 hectares)

future_PIAL <- rasterToPolygons(pial.surf_future, function(x) { x > 0.7 },dissolve=TRUE)

sum(sapply(slot(future_PIAL, "polygons"), slot, "area"))*0.0001
```

That's a change of 975.33 hectares in area with a 1 deg warming.

Assignment 7 (Optional)

Your goal this week is to use a neural net classifier to classify an image and report the overall accuracy, kappa coefficient, and confusion matrix of the classification.

Requirements:

1) The function should be named “machine_learning_classification” and have the following parameters (no defaults):

x: the input multispectral file

training: a SpatialPointsDataFrame with known SPECIES
to be used to train the classifier.

testing: a SpatialPointsDataFrame with known SPECIES
to be used to perform accuracy assessment.

Assignment 7 is due on April 30, 2020

Assignment 7 (Optional)

2) The function should extract the 3-bands of data at the training and testing locations.

3) The function nnet(), which is part of the package nnet (you will need to install it) should be used to create the classifier based on the training data. All function parameters should be left to their default values except:

number of units in the hidden layer: 2

initial random weights on [-rang,rang]: 0.1

weight decay parameter: 5e-4

maximum number of iterations: 1000

Assignment 7 (Optional)

4) Use the output model to predict the SPECIES from the testing dataset pixel values.

5) A confusion matrix and accuracy stats should be generated using the nnet predicted SPECIES and the testing SPECIES. Use the confusionMatrix function in package 'caret'.

6) The nnet model is applied to the input raster (1 point of extra credit).

7) The function should return a list with two elements named "confusion_matrix" and "classified_image”, which are the confusionMatrix() function output, and the classified raster (for the extra credit).

8) Comment your code in at least 3 places.

9) The code should be submitted tom Compass 2g as a single function with the filename: Lastname-Firstname-geog489-s20-assignment7.R