Exercise 6: Predictors - Automated vs Thoughtful Selection

Adam B. Smith, Danielle Svehla, & Camilo Sanín

July 19, 2017

The 19 BIOCLIM variables provide a helpful way to think about climate. However they are not the only variables one could use and they do have drawbacks and deficiencies. In the past, it was common to see studies in which all 19 variables were used in modeling. Even today you can still encounter publications that do so. These models might create valid-looking output, but they may not be the best models for projecting to the future or other regions. This can occur because environmental variables are often correlated with one another, so a less important variable that is correlated with several influential variables might be chosen by the model because doing so creates a more parsimonious model.

There is a growing debate about whether thoughtful and informed variable choice is better or letting the "machine" do the choosing is better. Until recently thoughtful choosing was preferred, though as noted people sometimes still used all available data. Some recent papers suggest that machine-assisted selection may do better, but a consensus hasn't been achieved. In this workshop we'll use user-chosen variables, but this first modeling exercise demonstrates the difference between the two approaches.

When choosing variables attempt to choose variables that are "proximate" (have a direct, mechanistic influence) on the species as opposed to "distal", indirect predictors that likely operate through other factors. For example, mean annual temperature and precipitation probably have little direct relation to a species since calculation of an average can mask relevant extremes. A species may very rarely experience the "mean" conditions of its environment. In contrast, extreme heat/cold can have direct physiological effects and drive energy requirements and resource availability, while extremes in moisture can drive resource availability.

Here we will:

1. **Train** our first models while
2. Exploring the difference between **machine-selected versus ecologist-selected variables**.

# Background sites

Ideally you would have data from surveys that indicated not only the presence but also the absence of a species. Unfortunately, absences are not usually recorded in biodiversity databases--and if they were, they'd be subject to the problem of false absence when a species was really there but not recorded. So even if you have "absences" it's important to consider whether they're useful.

If you had reliable absences you could use a generalized linear model (GLM) or something similar to predict the actual probability of presence. However, our data sets do not have absences. As a result, we'll use **background sites** which are points used to represent the environment across the landscape of interest. Background sites can fall on known presence and absence sites--this is legitimate. Initially we'll use background sites randomly placed across the study region.

First, we'll create our predictor stack of rasters.

climate <- stack(c(  
 './WORLDCLIM/Elevation/Study Region/elevation.tif',  
 list.files('./WORLDCLIM/1970-2000/Study Region',  
 full.names=TRUE)  
))

Now let's randomly locate background sites.

# generate 10,000 background sites  
randomBgSites <- randomPoints(climate, 10000)  
  
# extract environment at sites  
randomBgEnv <- extract(climate, randomBgSites)  
randomBgEnv <- as.data.frame(randomBgEnv)  
  
# remove any sites with NA for at least one variable  
isNa <- is.na(rowSums(randomBgEnv))  
if (any(isNa)) {  
 randomBgSites <- randomBgSites[-which(isNa), ]  
 randomBgEnv <- randomBgEnv[-which(isNa), ]  
}  
  
# combine with coordinates and rename coordinate fields  
randomBg <- cbind(randomBgSites, randomBgEnv)  
names(randomBg)[1:2] <- c('longitude', 'latitude')  
head(randomBg)

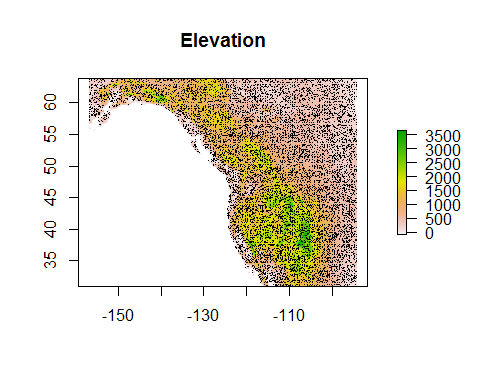
## longitude latitude elevation WC01 WC02 WC03 WC04 WC05 WC06 WC07 WC08  
## 1 -108.7500 31.91667 1404 16 20 49 756 35 -5 40 24  
## 2 -102.4167 39.25001 1306 10 16 38 940 31 -10 42 19  
## 3 -139.9167 62.08334 804 -4 11 24 1367 18 -30 48 12  
## 4 -99.2500 33.58334 408 17 14 37 873 36 -2 38 23  
## 5 -147.4167 61.91667 1181 -3 8 25 965 14 -20 34 9  
## 6 -109.4167 41.08334 2120 6 16 39 925 26 -15 41 9  
## WC09 WC10 WC11 WC12 WC13 WC14 WC15 WC16 WC17 WC18 WC19  
## 1 19 25 6 320 66 5 72 167 22 137 70  
## 2 -1 22 -1 407 74 8 68 199 27 178 27  
## 3 -11 12 -21 417 86 14 67 212 44 212 48  
## 4 7 28 6 679 102 24 45 233 89 192 91  
## 5 -10 9 -14 450 68 15 44 185 66 185 82  
## 6 -6 17 -6 271 34 12 30 94 40 78 40

Let's save these for later use.

dir.create('./Background Sites', recursive=TRUE, showWarnings=FALSE)  
save(randomBg, file=  
 './Background Sites/Random Background Sites across Study Region.Rdata',  
 compress=TRUE)

Take a look!

plot(climate[['elevation']], main='Elevation') # plot first raster in climate stack  
points(randomBgSites, pch='.')



# A model with automated variable selection

To train our model we'll need a data frame with the environmental data from the presence sites and the background sites. This data frame should *only* contain the variables we want the model to use, which in our case is WC01 through WC19. By including all of these in the data frame we're allowing the model algorithm to choose which variables to use.

trainDataAuto <- rbind(  
 records[ , c(paste0('WC0', 1:9), paste0('WC', 10:19))],  
 randomBgEnv[ , c(paste0('WC0', 1:9), paste0('WC', 10:19))]  
)

We also need a vector of 1's and 0's, one value per row in trainData to indicate if it's a presence or a background site.

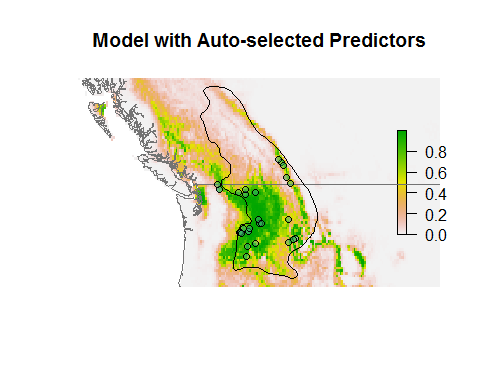
presBg <- c(rep(1, nrow(records)), rep(0, nrow(randomBgEnv)))

Let's train our first Maxent model!

# model species: note we're using the new "maxnet" function, not the older "maxent" function!  
autoSelectModel <- maxnet(p=presBg, data=trainDataAuto)  
  
# create output directory... one per model  
dir.create('./Models/Model 01 Predictors - Automated Selection', recursive=TRUE,  
 showWarnings=FALSE)  
  
# save model  
save(autoSelectModel,  
 file='./Models/Model 01 Predictors - Automated Selection/Model.Rdata',  
 compress=TRUE)

Later we will spend time on evaluating the model's performance, but for now we'll just focus on predictor selection. Let's write the model output as a raster prediction then view it.

# save raster  
autoSelectMap <- predict(climate, autoSelectModel,  
 filename='./Models/Model 01 Predictors - Automated Selection/maxentPrediction1970to2000',  
 format='GTiff', overwrite=TRUE, type='cloglog')  
  
# plot output  
plot(rangeMap, main='Model with Auto-selected Predictors')  
plot(autoSelectMap, add=TRUE)  
sp::plot(countries, add=TRUE, border='gray45')  
plot(rangeMap, add=TRUE)  
points(records$longitude, records$latitude, pch=21, bg=alpha('mediumseagreen', 0.5))



What do you see? *Recall that the range map is probably drawn at a coarse scale and ultimately derived from specimen data itself*, so using it as a visual benchmark is circular. However, it can give us a general idea of how much the model over- and underpredicts. The model mainly predicts that areas in the US are suitable, whereas areas in Canada are not. This is probably a reflection of the lack of data from Canada. In subsequent tutorials we'll explore how to correct for this.

Also note that the model does predict that the Coast Mountains in southwestern British Columbia are suitable, even though there are no presences there and the range map suggests the species does not occur there. Depending on your goals, this may or may not be an error. If you're modeling the distribution, it's an overprediction. But if you're modeling the geographic representation of the niche, it may not be--the species could actually live there, but perhaps it doesn't due to dispersal limitations or competition, disease, or predators.

# Informed variable selection

Now let's compare our model with automatically-selected variables with one using more thoughtfully-selected variables. Consider the list of BIOCLIM variables (see Exercise 2 or search the Internet for "BIOCLIM" and "WORLDCLIM"--it will be the first result). Which is these are likely to impose direct physiological or resource limits on our species?

We will conduct this process by choosing a set of "candidate" predictors which we then cull based on correlations between them. Alternatively, we could condense our candidates into a set of principal component axes and use the first several axes as predictors. The advantage of the latter approach is that it ensures predictors are uncorrelated. The disadvantage is that predictors are more difficult to interpret. We'll stick with using the predictors as-is.

We'll choose candidate predictors on the basis of some very general biology for each species. We'll *not* use mean annual temperature (WC01) and mean annual precipitation (WC12). Although these variables are easy to comprehend, they probably have little direct bearing on species' biology since they reflect central tendency. To locations could have the same mean annual temperature, for example, but one could be very steady and anther very variable in its temperature profile.

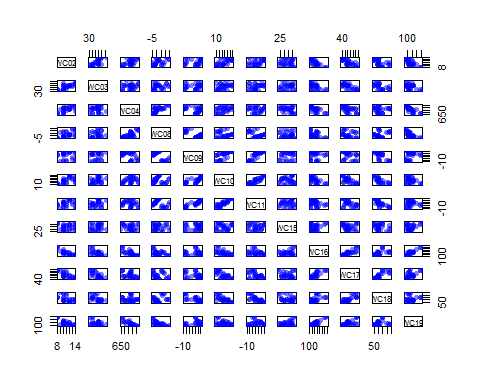
The Columbian Ground Squirrel is a superb hibernator--in fact, it can spend the majority of the year in hibernation, emerging ravenously hungry. Hence, it's probably buffered from short-term climatic variability. However, the plants on which it depends must be able to grow enough to supply the squirrel with energy. So on this very general basis, let's choose candidate predictors that reflect seasonal (versus monthly) values of temperature and precipitation. We'll also choose variables that reflect annual variabiliy in temperature or precipitation since these likely reflect the difficulty of the squirrels and plants on which they depend to adapt.

candidates <- c('WC02', 'WC03', 'WC04', 'WC08', 'WC09', 'WC10', 'WC11', 'WC15', 'WC16', 'WC17', 'WC18', 'WC19')

## Correlations between variables

Now, let's examine correlations between the variables. Our goal is to choose among highly-correlated variables so that in the end none have string relationships. Here, "strong" is often taken to mean a correlation coefficient with an absolute value >~0.7 or so. Let's examine the correlations between variables.

pairs(records[ , candidates], col=alpha('blue', 0.5), pch=16)



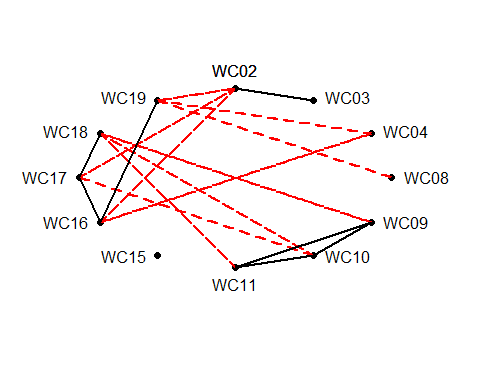
Some variables are highly correlated! Let's look at this in tabular form. We'll use the Spearman rank correlation because the distribution of variables is often very non-normal.

correl <- cor(records[ , candidates], method='spearman')  
print(correl, digits=2)

## WC02 WC03 WC04 WC08 WC09 WC10 WC11 WC15 WC16 WC17  
## WC02 1.000 0.75 0.635 0.57 0.26 0.604 0.315 -0.038 -0.820 -0.77  
## WC03 0.747 1.00 0.120 0.23 0.31 0.379 0.364 0.130 -0.537 -0.65  
## WC04 0.635 0.12 1.000 0.68 -0.11 0.365 -0.055 -0.202 -0.743 -0.53  
## WC08 0.568 0.23 0.682 1.00 -0.31 0.138 -0.175 -0.138 -0.462 -0.35  
## WC09 0.263 0.31 -0.107 -0.31 1.00 0.841 0.957 0.221 -0.255 -0.51  
## WC10 0.604 0.38 0.365 0.14 0.84 1.000 0.891 -0.014 -0.601 -0.70  
## WC11 0.315 0.36 -0.055 -0.17 0.96 0.891 1.000 0.121 -0.325 -0.55  
## WC15 -0.038 0.13 -0.202 -0.14 0.22 -0.014 0.121 1.000 0.091 -0.27  
## WC16 -0.820 -0.54 -0.743 -0.46 -0.25 -0.601 -0.325 0.091 1.000 0.86  
## WC17 -0.766 -0.65 -0.531 -0.35 -0.51 -0.700 -0.552 -0.267 0.863 1.00  
## WC18 -0.477 -0.58 -0.090 0.12 -0.84 -0.780 -0.847 -0.205 0.576 0.80  
## WC19 -0.727 -0.33 -0.864 -0.74 0.13 -0.311 0.045 0.140 0.843 0.66  
## WC18 WC19  
## WC02 -0.48 -0.727  
## WC03 -0.58 -0.326  
## WC04 -0.09 -0.864  
## WC08 0.12 -0.744  
## WC09 -0.84 0.130  
## WC10 -0.78 -0.311  
## WC11 -0.85 0.045  
## WC15 -0.20 0.140  
## WC16 0.58 0.843  
## WC17 0.80 0.659  
## WC18 1.00 0.193  
## WC19 0.19 1.000

We can see that some variables have correlations >0.70. With so many variables the table is difficult to interpret, so let's use another tool to see the same thing. Since we're only interested in correlations >0.7, let's focus on those. We've written a script to create a "spoke plot" that draws lines between variable names if they are correlated more than a given amount.

pos <- correl > 0.7  
neg <- correl < -0.7  
  
spoke(  
 pos=pos,  
 neg=neg,  
 lwdPos=2,  
 lwdNeg=2,  
 colPos='black',  
 colNeg='red',  
 pty='s'  
)



Here black lines join variables that are highly positively correlated (>0.7) and red lines variables that are highly negatively correlated (<-0.7). It doesn't matter if variables are highly negatively or positively correlated--having information on one variable gives you information about the other, so we need to eliminate statistically redundant variables.

*How many* should we choose? There's no set rule, but simple ecological models suggest that at most a species' range could be limited by a handful (roughly to 5-7) orthogonal variables before there would simply be no place to live ([Jim Brown. 1995. Macroecology. U Chicago Press](http://press.uchicago.edu/ucp/books/book/chicago/M/bo3632297.html)).

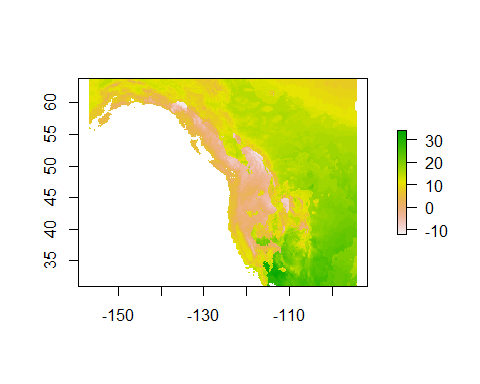
So *how* should we choose? The candidate variables that are not strongly correlated with anything else should be retained. Regarding correlated variables, keep in mind that if you have to choose between them when both seem reasonable choosing one still represents the information content of the other to a degree. Personally, when choosing among variables we try to keep the ones that are easier to explain (e.g., minimum temperature versus variability in temperature). Given these general how-to statements, let's choose based on species-specific reasoning.

Please look at the spoke plot we created above and keep it in mind as we think through variable selection. First, note that WC15 (precipitation variability) is uncorrelated with anything else, so let's retain it.

Now, note that WC02, WC03, and WC04 each reflect some aspect of annual temperature variation. These variables are often important in modeling exercises, but it's also more difficult to posit direction relationships between them and the persistence of species than more "direct" variables like winter or summer temperature. So let's mentally put WC02 through WC04 aside for now and select some representatives of both temperature-related variables (WC01 through WC11) and precipitation-related variables (WC12 through WC19).

First, note that some of the variables in this list are "crossed" variables, meaning they reflect *x* relative to a period defined by *y*. For example, WC08 is mean temperature of the wettest quarter. (A quarter can be any consecutive 3 months--June, July, August, or December, January, February). WC08 sounds like an interesting variable since it reflects how much heat energy is available during the wettest period. However, the wettest quarter can change a lot across some landscapes, meaning this particular temperature variable can reflect very different times of the year. To see this, let's plot it:

plot(climate[['WC08']])



It seems this variable reflects winter temperature along the coast and summer temperature further inland. WC09, mean temperature of the driest quarter, is similar. So let's not use WC08 or WC09. The same reasoning applies to WC16 and 17, precipitation of the wettest and driest quarters. (Note that we could have simply not chosen them as candidates in the first place... but sometimes it's informative to see something done badly.)

So this leaves WC10, WC11, WC18 and WC19. WC10 and 11, summer and winter temperature, are correlated with one another, so in theory we could arbitrarily choose one. But recall that the Columbian ground squirrel hibernates, so it's probably not experiencing much of WC11. Let's stick with WC10.

WC10 is correlated with WC18. So we could choose WC18 instead and drop WC10, but it seems unlikely that the species responds to *just* precipitation, so let's keep WC10 and drop WC18. That leaves WC19 (winter precipitation).

Now, let's go back to the "temperature variation" variables, WC02, WC03, and WC04. Note that we chose WC19, which is highly correlated with WC02, so let's not use WC02. WC19 is also correlated with WC04, so let's also discard WC04. WC03 is only highly correlated with WC02 (which we're not using), so let's keep WC03.

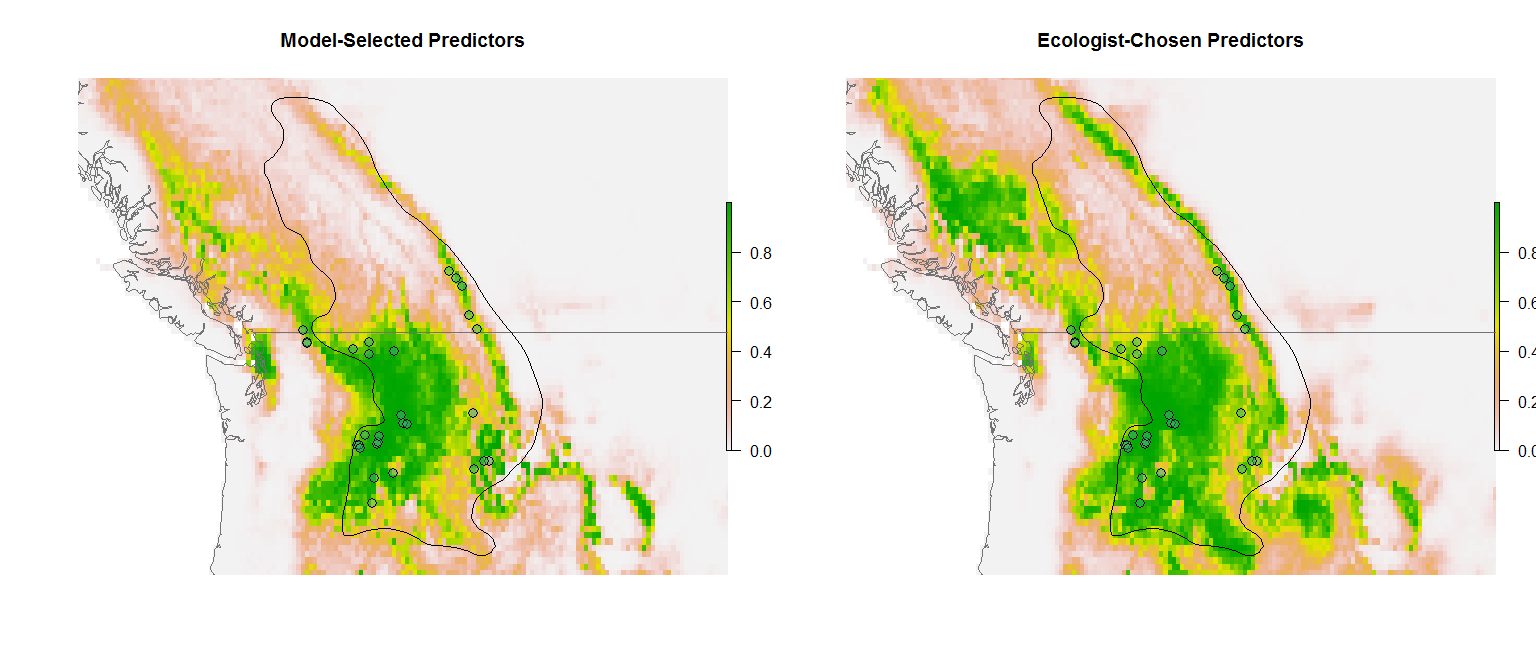
This process left us with just four variables: WC03 (isothermality--the range of annual temperature divided by the mean range of monthly temperature, WC10 (summer temperature), WC15, (annual variation in precipitation), and WC19 (winter precipitation). Is this enough? We'll see... machine-automated selection is looking a little easier...

predictors <- c('WC03', 'WC10', 'WC15', 'WC19')

## Modeling with thoughtfully-selected variables

We'll use the same set of random background sites used in the previous model. We'll also use the same presences, of course!

# select chosen predictor variables  
trainDataManual <- rbind(records[ , predictors], randomBg[ , predictors])  
  
# vector of 1/0 to indicate presence/background  
presBg <- c(rep(1, nrow(records)), rep(0, nrow(randomBg)))  
  
# output directory  
dir.create('./Models/Model 02 Predictors - Manual Selection',  
 recursive=TRUE, showWarnings=FALSE)  
  
# model  
manualSelectModel <- maxnet(p=presBg, data=trainDataManual)  
  
# save model  
save(manualSelectModel,  
 file='./Models/Model 02 Predictors - Manual Selection/Model.Rdata',  
 compress=TRUE)  
  
# write raster  
climateSelect <- subset(climate, c('WC03', 'WC10', 'WC15', 'WC19')) # speeds things up to use just the necessary rasters  
manualSelectMap <- predict(climateSelect, manualSelectModel,  
 filename='./Models/Model 02 Predictors - Manual Selection/maxentPrediction1970to2000',  
 format='GTiff', overwrite=TRUE, type='cloglog')  
  
# plot  
par(mfrow=c(1, 2))  
plot(rangeMap, main='Model-Selected Predictors')   
plot(autoSelectMap, add=TRUE)  
sp::plot(countries, add=TRUE, border='gray45')  
plot(rangeMap, add=TRUE)  
points(records$longitude, records$latitude, pch=21, bg=alpha('mediumseagreen', 0.5), cex=1.4)  
  
plot(rangeMap, main='Ecologist-Chosen Predictors')  
plot(manualSelectMap, add=TRUE)  
sp::plot(countries, add=TRUE, border='gray45')  
plot(rangeMap, add=TRUE)  
points(records$longitude, records$latitude, pch=21, bg=alpha('mediumseagreen', 0.5), cex=1.4)



So what's going on here? With just 4 variables the prediction is more generous--maybe even too much so! Why is this? When the model has just a few variables, it has less ability to say "no" to a high prediction--after all, it has very limited information and so any place that has values of WC03, WC10, WC15, and WC19 similar to those where there are presences is, in the model's mind, a suitable place. On the other hand, by using all 19 variables we have a lot of ways to say "no"--any one of those variables can be different from the presences. *However* using all 19 variables might be too much--after all, even a random variable has some explanatory power.

# Reflection

1. Which model looks to be a better depiction of the species' range? Why?
2. For a given species in which you are interested, what variables do you think would have the most proximate (direct) effect on the distribution of the species? Is there empirical work to suggest which factors most influence the species?
3. Do you think the [19 BIOCLIM variables](http://www.worldclim.org/bioclim) adequately represent the full range of climatic variables that drive species' distributions? What's missing? Do the 19 BIOCLIM variables better reflect factors that influence distributions of temperate, tropical, or boreal/arctic species?
4. Under what circumstances is it appropriate to choose "crossed" variables like WC08 and WC09 (temperature of the wettest or driest quarter) or WC18 or WC19 (precipitation of the warmest and coldest quarter)?