Exercise 5: Match Records to Environment  
and Scrutinize for Outliers

Adam B. Smith, Danielle Svehla, & Camilo Sanín

July 19, 2017

We've already scrutinized our species' data for erroneous records in geographic space. However, what may appear legitimate on a map may not actually be legitimate. Consider a species that lives in the lowlands. It may have a record that falls inside its range when viewed on our computer screen, but on closer inspection the point actually falls on a mountaintop. These are "pernicious" errors that cannot be easily avoided. One way to help catch them (and geographic outliers) is to examine the species' distribution in environmental space.

# Matching environmental data to records

We will create a raster stack which stores one or more rasters and can be manipulated like a single variable. We'll then extract the raster values at each record location.

env <- stack(c(  
 './WORLDCLIM/Elevation/Study Region/elevation.tif',  
 list.files('./WORLDCLIM/1970-2000/Study Region',  
 full.names=TRUE)  
))  
  
# match species' records with environment at each location  
envSpecies <- extract(env, cbind(records$longitude, records$latitude))  
envSpecies <- as.data.frame(envSpecies)  
records <- cbind(records, envSpecies)  
head(records)

## dataSet idNum rawSpecies longitude  
## 889 GBIF 1145723365 Urocitellus columbianus (Ord, 1815) -114.0349  
## 890 GBIF 1145723276 Urocitellus columbianus (Ord, 1815) -114.0349  
## 932 GBIF 1145140596 Urocitellus columbianus (Ord, 1815) -113.8654  
## 933 GBIF 1145140565 Urocitellus columbianus (Ord, 1815) -113.8654  
## 934 GBIF 1145130699 Urocitellus columbianus (Ord, 1815) -116.8226  
## 935 GBIF 1145130698 Urocitellus columbianus (Ord, 1815) -116.8921  
## latitude coordUncer recordType country state county  
## 889 46.87199 9298.79 PhysicalObject US Montana Missoula County  
## 890 46.87199 9298.79 PhysicalObject US Montana Missoula County  
## 932 49.08806 50.00 PhysicalObject CA Alberta   
## 933 49.08806 50.00 PhysicalObject CA Alberta   
## 934 46.62181 1142.63 PhysicalObject US Idaho Latah County  
## 935 46.84138 579.36 PhysicalObject US Idaho Latah County  
## locality  
## 889 Grant Creek Ranch, 2 mi W Missoula  
## 890 Grant Creek Ranch, 2 mi W Missoula  
## 932 Waterton Lakes National Park, about 3 miles NE of Waterton Park off Route 5  
## 933 Waterton Lakes National Park, about 3 miles NE of Waterton Park off Route 5  
## 934 7.5 miles SE of Tomer Butte  
## 935 2.88 miles NNW of Moscow Mountain  
## date year institution identifiedBy elevation WC01  
## 889 1975-06-18T01:00Z 1975 MVZ James L. Patton 1056 7  
## 890 1975-06-18T01:00Z 1975 MVZ James L. Patton 1056 7  
## 932 1988-08-01T02:00Z 1988 MSB Jonathan L. Dunnum 1670 3  
## 933 1988-08-01T02:00Z 1988 MSB Jonathan L. Dunnum 1670 3  
## 934 1975-07-13T01:00Z 1975 MSB Jonathan L. Dunnum 669 9  
## 935 1975-05-29T01:00Z 1975 MSB Jonathan L. Dunnum 845 8  
## WC02 WC03 WC04 WC05 WC06 WC07 WC08 WC09 WC10 WC11 WC12 WC13 WC14 WC15  
## 889 13 36 855 24 -12 36 15 3 18 -3 357 46 20 27  
## 890 13 36 855 24 -12 36 15 3 18 -3 357 46 20 27  
## 932 10 33 793 18 -14 32 6 -7 12 -7 832 105 53 23  
## 933 10 33 793 18 -14 32 6 -7 12 -7 832 105 53 23  
## 934 12 36 764 25 -9 33 1 19 19 0 596 69 26 30  
## 935 11 36 736 22 -9 32 0 17 17 -1 652 83 24 37  
## WC16 WC17 WC18 WC19  
## 889 119 71 100 85  
## 890 119 71 100 85  
## 932 275 166 221 166  
## 933 275 166 221 166  
## 934 201 84 96 186  
## 935 243 86 97 226

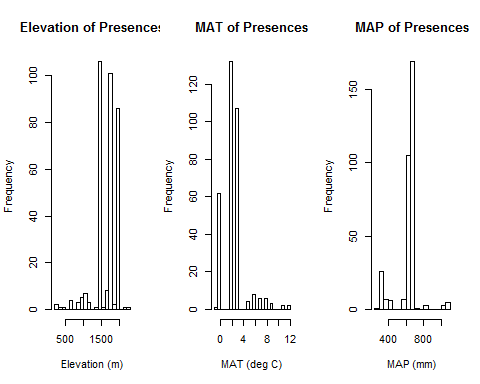
Sometimes you'll find a record that falls in a place where it should not (e.g., a terrestrial species in the water). This may be an error, but sometimes not... Raster data represents the world with squares, and a record may have legitimately been made in the "corner" of a cell that was in a valid habitat yet falls into an invalid habitat in the raster data set. For our purposes, let's remove any records that might have NA's as environmental data (i.e., fall into the water). (In our case none do, but it's good to check.)

if (any(is.na(rowSums(envSpecies)))) records <-  
 records[-which(is.na(rowSums(envSpecies))), ]

## Outliers in 1-D environmental space

Now, let's examine the distribution of records in 1- and 2-dimensional environmental space. Anything that seems excessive should be scrutinized. There are many variables we could use to examine environmental location. Here we'll use elevation plus mean annual temperature (MAT) and precipitation (MAP) since they represent "gestalt" variables.

# plot in 1-D environmental space  
par(mfrow=c(1, 3))  
hist(records$elevation, breaks=20, xlab='Elevation (m)', main='Elevation of Presences')  
hist(records$WC01, breaks=20, xlab='MAT (deg C)', main='MAT of Presences')  
hist(records$WC12, breaks=20, xlab='MAP (mm)', main='MAP of Presences')



If you have a lot of records, you can simply remove those that are, say, in the outermost 0.5% quantiles or farther than, say, 3 SD from the next-nearest point. Beware, though that this could exclude legitimate records that tell you a lot about the species' environmental tolerances.

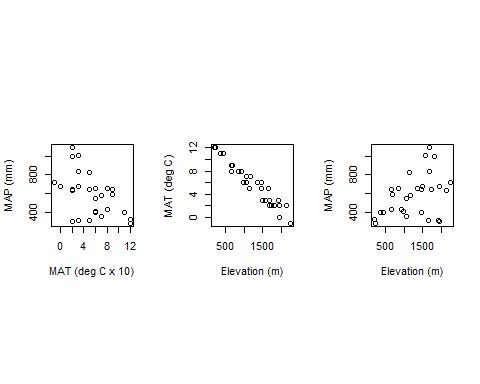
Note that there are a relatively large number of records with the same values of elevation, MAT, and MAP. This might reflect either highly suitable climate/elevation OR an overconcentration of collection effort in those locations. We'll examine this later.

Overall, though, there don't seem to be any obvious outliers. However, if there were, you should: \* Locate those records on a map. Do they represent likely habitat? \* Examine the data associated with the records. Do they have large coordinate uncertainty? Were they collected under odd circumstances? Were they correctly identified? Do any notes indicate something informative (e.g., caught in a sheltered cove? georeferenced to the zoo the animal was living in)?

## Outliers in 2-D environmental space

Sometimes a record can fall within the range of each variable and yet be an outlier in 2- or higher-dimensional environmental space. Let's look for these records.

par(mfrow=c(1, 3), pty='s')  
plot(records$WC01,  
 records$WC12,  
 xlab='MAT (deg C x 10)',  
 ylab='MAP (mm)',  
 main=''  
)  
  
plot(records$elevation,  
 records$WC01,  
 xlab='Elevation (m)',  
 ylab='MAT (deg C)',  
 main=''  
)  
  
plot(records$elevation,  
 records$WC12,  
 xlab='Elevation (m)',  
 ylab='MAP (mm)',  
 main=''  
)



Again, there don't seem to be obvious outliers. Let's save the records with their associated environmental values.

saveRDS(records, './Species Records/04 Species Records - Matched with WORLCLIM Data.rds')

# Remove all but one record per cell

Our record data has many duplicates in cells in the environmental rasters. Duplicates occur when a collector obtains more than one specimen at a site. They also occur when more than one site in a cell was surveyed. Having more than one record per cell isn't necessarily bad--it might actually reflect favorable environment. Nevertheless, it is standard practice to remove all but one record per cell to reduce the effects of sampling bias on models.

Let's remove all but one record per cell. The enmSdm package contains a function elimCellDups().

# use elevation raster as a "template" for finding cell duplicates  
elevation <- raster('./WORLDCLIM/Elevation/Study Region/elevation.tif')  
recordsNoDups <- elimCellDups(x=records, r=elevation, longLat=c('longitude', 'latitude'))  
  
nrow(records)

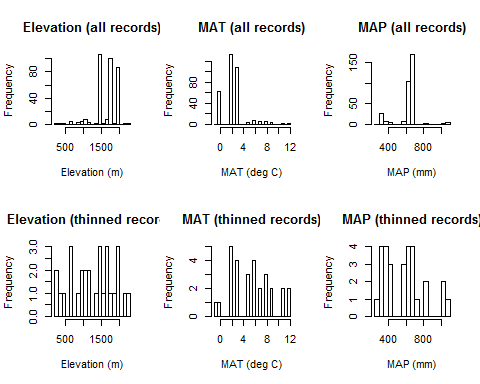
## [1] 333

nrow(recordsNoDups)

## [1] 29

How many records do we have left? (Recall we started by downloading 1000s of records!) What effect does this have on the distribution of environmental variables among records?

par(mfrow=c(2, 3))  
hist(records$elevation, breaks=20, xlab='Elevation (m)', main='Elevation (all records)')  
hist(records$WC01, breaks=20, xlab='MAT (deg C)', main='MAT (all records)')  
hist(records$WC12, breaks=20, xlab='MAP (mm)', main='MAP (all records)')  
  
hist(recordsNoDups$elevation, breaks=20, xlab='Elevation (m)', main='Elevation (thinned records)')  
hist(recordsNoDups$WC01, breaks=20, xlab='MAT (deg C)', main='MAT (thinned records)')  
hist(recordsNoDups$WC12, breaks=20, xlab='MAP (mm)', main='MAP (thinned records)')



Finally, let's save the thinned records to use for modeling.

records <- recordsNoDups  
saveRDS(records, './Species Records/05 Species Records - Eliminated All But One Record Per Cell.rds')

# Reflection

1. What methods can you use to identify outliers in geographic and environmental space? How can you use one kind of space to identify outliers in the other?
2. What effect did removing cell outliers have on the distribution of environmental variables across the species' presences? What effect do you think this will have on models?
3. We removed duplicates in each cell, but should we have done this? What do duplicates represent, better habitat or more sampling effort? How could you tell? (Note that we don't actually know the answers to this--it will depend on your situation, and the answers may be unknowable.)