Social Media Presence and Derivative Indices for Sampling Bias

Gayan Seneviratna

February 17, 2019

#Required Packages  
library("sdm")  
library("knitr")  
library("raster")  
library("sp")  
  
#Helper Functions  
source("Helpers/Grid2Raster.R")  
  
opts\_chunk$set(cache=TRUE)

The following report details my attempts to build an index of “human bias” in citizen science biodiversity data.

# The Issue

The recent popularity of citizen science has been a boon to biodiversity researchers, who now have access to massive collections of biological records like the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia (ALA). This has been especially useful for computational biologists in the development of SDMs (species distribution models). These models apply machine learning techniques on spatial data (typically ecological variables like precipitation or soil substrates) and use them to build predictive maps of species’ spatial ranges. l predictors for the occurrence of a given taxa. In doing so, the an use recorded data to build maps of species locations that can be used for conservation direction.

However, it has been well-established that this data is imperfect. Firstly, there are often a number of incorrect details recorded in citizen science data- misidentified species or impossible geography for example. Many of these issues can be identified and possibly corrected using a number of methods, as detailed in in Gueta, 2016.

However, even clean data can be imperfect due to bias. Bias is defined as consistent pattern in sampled data (or models built on them) that deviates from the true values. In the case of volunteer-provided data, there are a number of sampling effort biases that may be at work, which reflect the reality of inexact data collection standards.

The problems in these collections methods are diverse. Kadmon, 2004 demonstrated that data on woody plants in Israel were significantly biased to be nearer to roads than not. This reflects the likelihood that volunteers do not (or perhaps cannot) travel far beyond established road networks. In a similar study, Miller, demonstrated the of a “cottage effect”, where volunteers were more likely to record at more visually-appealing locations (i.e. where a cottage might be built) (2017). All of these effects are exacerbated by the fact that most volunteer collected data is presence-only, meaning that volunteers do not typically record instances of not finding a species. As such, it is difficult to determine whether a lack of data in a region is due to a lack of species or of human volunteers.

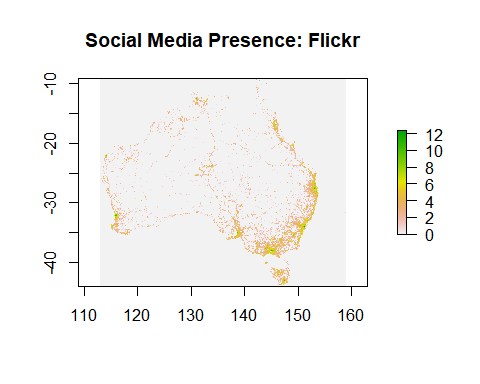
Monitoring these biases appears simple. We simply count the distance to a road to gauge the roadside bias. Pulling housing price data is a decent proxy for the cottage effect. However, this ends up being a limited view of the bias. Firstly, there are a large number of potential factors at work, beyond those explored in recent studies. Furthermore, there is likely to be an interaction of these factors- a bias of biases. Attempting to consider all of these covariates at once is going to become prohibitively complicated for many researchers.

Instead, one could attempt to directly capture the sampling effort. This is the logic behind ignorance maps. Using a reference species (or a higher taxon) which researchers expect is subject to the same bias but is more heavily sampled than the species of interest, researchers can dictate that various grid cells are more likely to be sampled overall. There are three major issues here. First, the identification of a reference species may be difficult. It would be simpler if there was a single measure that could be applied to any data set that a researcher believes might be biased. Second, there may be variance in the ignorance scores that is relevant for predictions. For example, we as humans tend to establish communities in fertile soils. That soil quality may also be a predictor for a species. In including ignorance as a predictor, we might end up ignoring the very real effect of soil fertility. Third, and perhaps most importantly, ignorance maps may not be subjected to the same biases, since volunteer efforts often center on a specific species..

# The Solution

To address these issues, this report presents two innovations. First, a new measure of sampling effort related directly to the probability of a person being at a given location: social-media presence. Social media-presence refers to the amount of activity tied to a given location by users of a social networking platform. The term is purposely left vague, as the choice of platform from which to identify activity may vary from locale to locale. In this use case, I am using posts from Flickr (a photo-sharing service), but this same experiment could have also been carried out with YouTube video posts or Facebook posts. By counting the amount of posts made from a given location unit, a map of social-media presence can be built:

#Load Australia 0.1 x 0.1 degree grid  
grid <- shapefile("Data/Reference/Grid/Grid\_0.1.shp")  
  
#Reads in and uses a custom function meant to turn grid data into a raster  
flickr <- read.csv("Data/Human/Flickr/flickr.csv", stringsAsFactors=FALSE)  
names(flickr) <- tolower(names(flickr))  
flickr <- grid2raster(flickr, log(flickr$photocount + 1))  
  
#Plots the flickr data  
sp::plot(flickr, main="Social Media Presence: Flickr", xlim = c(112, 159))



A number of trends are immediately evident in this plot. First, there is a large concentration of activity around major cities like Sydney and Melbourne. Unlike simple indicators of city boundaries, this map shows a more fine grained look at activity. At first glance then, social media presence is no different from population maps.

However, there are also spots of activity in areas without high populations. The collection of green cells located at the center are both Alice Springs and national parks Watarrka and West MacDonnell; areas of low living population but high foot traffic. There are also a number of paths along major routes in Australia (with higher density than would be expected for roadside villages). This reflects the ability of this data to capture multiple biases in a single variable. The fact that all of these trends are present in a single variable suggests that social media presence is a measure of “where people go” rather than “where people live”.

The source of this data (online queries of all posts on a given platform) also presents a number of advantages). First, they can provide a complete measure of all activity on the platform; there is no sampling bias since this the population of data. Second, the flexibility of many social media APIs means that the scale of analysis can also be adjusted as necessary. This analysis was made on the 0.1 x 0.1 degree scale, but a more fine grained approach can be done for exploring biases like the Cottage Effect. Third, the popularity of these services worldwide suggests that there would be good coverage for most regions that have citizen science work.

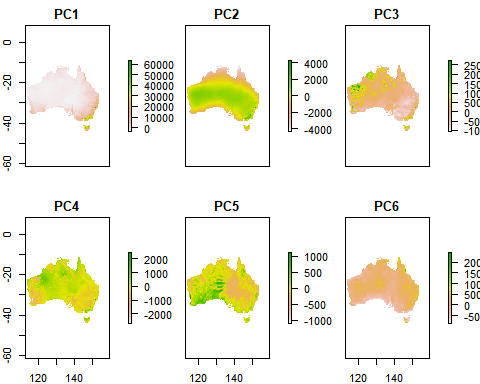
Before we explore the set of indices available using social media presence, it is also worthwhile to consider how this data relates to the natural predictors. Humans are species after all; their choice of habitat is also highly dependent on environmental factors. The following is the set of biological predictors used in this effort

load("Analytics/NaturalBrick.RData")  
  
predictors <- data.frame(Name = names(naturalBrick),  
 Source = c(rep("Bioclim", 25),  
 rep("CSIRO", 37),  
 rep("ALA", 2)))  
kable(predictors)

|  |  |
| --- | --- |
| Name | Source |
| alt | Bioclim |
| biome | Bioclim |
| ndvi | Bioclim |
| landuse | Bioclim |
| solar\_ann | Bioclim |
| evap\_ann | Bioclim |
| bio\_1 | Bioclim |
| bio\_2 | Bioclim |
| bio\_3 | Bioclim |
| bio\_4 | Bioclim |
| bio\_5 | Bioclim |
| bio\_6 | Bioclim |
| bio\_7 | Bioclim |
| bio\_8 | Bioclim |
| bio\_9 | Bioclim |
| bio\_10 | Bioclim |
| bio\_11 | Bioclim |
| bio\_12 | Bioclim |
| bio\_13 | Bioclim |
| bio\_14 | Bioclim |
| bio\_15 | Bioclim |
| bio\_16 | Bioclim |
| bio\_17 | Bioclim |
| bio\_18 | Bioclim |
| bio\_19 | Bioclim |
| bdensity | CSIRO |
| calcrete0 | CSIRO |
| clay | CSIRO |
| coarse0 | CSIRO |
| corg0 | CSIRO |
| distanywater | CSIRO |
| distnonpermw | CSIRO |
| distpermwat | CSIRO |
| erosional | CSIRO |
| fert | CSIRO |
| geollmeanage | CSIRO |
| geollrngeage | CSIRO |
| geolmeanage | CSIRO |
| geolrngeage | CSIRO |
| gravity | CSIRO |
| hpedality | CSIRO |
| ksat0 | CSIRO |
| magnetics | CSIRO |
| mrrtf | CSIRO |
| mrvbf | CSIRO |
| nmnlconcn0 | CSIRO |
| nmnln0 | CSIRO |
| ntotn0 | CSIRO |
| nutrients | CSIRO |
| pedality | CSIRO |
| pmnlconcn0 | CSIRO |
| pmnln0 | CSIRO |
| ptotn0 | CSIRO |
| relief | CSIRO |
| ridgetopflat | CSIRO |
| roughness | CSIRO |
| slope | CSIRO |
| soldepth0 | CSIRO |
| solpawhc0 | CSIRO |
| twi | CSIRO |
| valleybottom | CSIRO |
| wr\_unr0 | CSIRO |
| acacia | ALA |
| eucalyptus | ALA |

The following raster images are the results of a primary components analysis that identified at least 90% of the variance in 64 biological predictor variables, reducing dimensionality down to 6 variables.

#NOTE: This code does not currently work due to a failure in compiling the RSToolbox library. A pre-built set of PCs is presented instead, as built on another machine.  
#naturalPCA <- rasterPCA(naturalBrick)  
#vars <- naturalPCA$model$sdev  
#cumVars <- cumsum(vars / sum(vars))  
#which(cumVars > 0.89)[1]  
#NaturalBrickPCA <- naturalPCA$map  
#NaturalBrickPCA <- brick(  
# subset(NaturalBrickPCA, 1, pull=TRUE),  
# subset(NaturalBrickPCA, 2, pull=TRUE),  
# subset(NaturalBrickPCA, 3, pull=TRUE),  
# subset(NaturalBrickPCA, 4, pull=TRUE),  
# subset(NaturalBrickPCA, 5, pull=TRUE),  
# subset(NaturalBrickPCA, 6, pull=TRUE)  
# )  
  
#Load the natural data predictors  
load("Analytics/NaturalBrickPCA.RData")  
sp::plot(NaturalBrickPCA)



I am unsure of the exact interpretation of all of these PCs. That being said, there does appear to be some correlation to human patterns, as there is a lot of data on the coastlines of these sets, where the majority of people in Australia live. It also appears that PC2 is broadly speaking a measure of latitude and the climactic changes that implies.

Before building the models, I randomly generated some flickr occurrence data from the rasters.

#Adjust the flickr data to have a small probability on 0 values, for the purpose of   
# random point generation  
flickr[which(flickr[1:length(flickr)] == 0)] <- 1e-99  
  
#Generate points from the flickr raster  
n<-10000  
occ = data.frame(Occurence = rep(1, n))  
flickrOcc <- SpatialPointsDataFrame(randomPoints(flickr, n, prob=TRUE),  
 data=occ,  
 proj4string=crs(NaturalBrickPCA))  
  
#Limit the generated points to those in the grid  
keepIndices <- over(x=as(flickrOcc, "SpatialPoints"), y=as(grid,"SpatialPolygons"))  
flickrOcc <- flickrOcc[which(!is.na(keepIndices)),]

The modelling results are as follows:

#Build the SDM data for modelling  
flickrSDMData <- sdm::sdmData(Occurence~.,  
 train=flickrOcc,  
 predictors=NaturalBrickPCA,  
 bg=list(n=length(flickrOcc),  
 method='gRandom',  
 remove=TRUE))

## Loading required package: gbm

## Loading required package: survival

## Loading required package: lattice

## Loading required package: splines

## Loading required package: parallel

## Loaded gbm 2.1.3

## Loading required package: tree

## Loading required package: mda

## Loading required package: class

## Loaded mda 0.4-10

## Loading required package: mgcv

## Loading required package: nlme

##   
## Attaching package: 'nlme'

## The following object is masked from 'package:raster':  
##   
## getData

## The following object is masked from 'package:dplyr':  
##   
## collapse

## This is mgcv 1.8-17. For overview type 'help("mgcv-package")'.

## Loading required package: glmnet

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-16

## Loading required package: earth

## Loading required package: plotmo

## Loading required package: plotrix

## Loading required package: TeachingDemos

## Loading required package: rJava

## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Loading required package: rpart

## Loading required package: kernlab

##   
## Attaching package: 'kernlab'

## The following objects are masked from 'package:raster':  
##   
## buffer, rotated

#Perform the modelling; only relatively fast methods used here due to kntir slowdown  
flickrModel <- sdm::sdm(Occurence ~.,  
 data=flickrSDMData,  
 methods=c("glm", "glmnet", "brt", "rpart"))   
  
#Model summary  
kable(sdm::getEvaluation(flickrModel, wtest="dep.test", stat="2"))

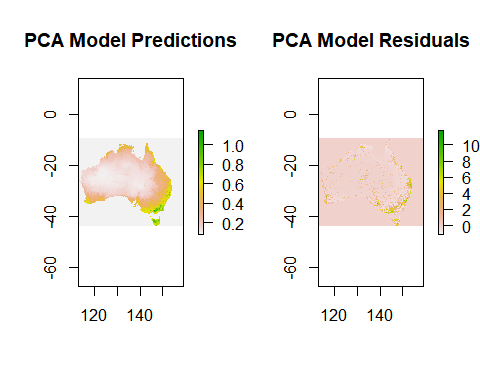
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| modelID | AUC | COR | Deviance | TSS |
| 1 | 0.783 | 0.491 | 1.131478 | 0.461 |
| 2 | 0.777 | 0.376 | 4.931297 | 0.462 |
| 3 | 0.766 | 0.457 | 1.215866 | 0.418 |
| 4 | 0.760 | 0.496 | 1.123934 | 0.450 |

According to these results, we can note that there is already a significant amount of variance in the social media presence that is explained by natural factors, accounting for about 45% of the total variance on average between the models. We can see what this looks like in terms of the predicted values. We also generate the residuals and will shortly provide a justification for their value.

#Generating predictions from an ensemble of the above models  
flickrPredictions <- ensemble(flickrModel, newdata=NaturalBrickPCA,setting=list(method='weighted',stat='AUC'))

## Warning in .rasterFromRasterFile(grdfile, band = band, objecttype, ...):  
## size of values file does not match the number of cells (given the data  
## type)  
  
## Warning in .rasterFromRasterFile(grdfile, band = band, objecttype, ...):  
## size of values file does not match the number of cells (given the data  
## type)

#Resample the predicted values to the same rate as the flickr data  
flickrPredictions <- resample(flickrPredictions, flickr)  
  
#Set the NA values to the minimum predicted value for plotting purposes (this will be removed for later use via cropping)  
flickrPredictions[which(is.na(flickrPredictions[1:length(flickrPredictions)]))] <- min(flickrPredictions[1:length(flickrPredictions)], na.rm=TRUE)  
  
#Generate the residuals  
flickrResiduals <- flickr - flickrPredictions  
  
#Plotting the results  
par(mfrow=c(1,2))  
plot(flickrPredictions, main="PCA Model Predictions")  
plot(flickrResiduals, main="PCA Model Residuals")



Note that the natural data only identifies some of the regions of human settlement, and even then appears to broadly map habitat areas rather than distinct population centers. It also appears to generally miss road-based effects. The residuals can be therefore interpreted as a measure of human activity, independent of the distribution of humans along ecological variables– a perhaps cleaner measure of bias that should not be in the data. As we will see shortly, this has the potential to be used in a number of measures of bias.

# The Indices

The second innovation we present are a set of procedures using social media presence, to summarize the degree of bias in a given region. By and large, these procedures are an extension of existing methods using the new predictor, with adjustments where interesting or relevant.

The data for these efforts are from GBIF, including occurrence records for four major mammals species in Australia: Kangaroos, Koalas, Flying Foxes, and Wombats.

#Index and grid data  
indexData <- grid  
  
#Read species data  
flyingFox <- shapefile("Data/Species/FlyingFox/FlyingFox.shp")  
kangaroo <- shapefile("Data/Species/Kangaroo/Kangaroo.shp")  
koala <- shapefile("Data/Species/Koala/Koala.shp")  
wombat <- shapefile("Data/Species/Wombat/Wombat.shp")  
  
#Downsample the more populous species to allow poly.counts to proceed efficiently  
kangaroo <- kangaroo[sample(1:length(kangaroo), 15000),]  
koala <- koala[sample(1:length(koala), 15000),]  
wombat <- wombat[sample(1:length(wombat), 15000),]  
  
#Add counts of mammal data  
indexData$FlyingFox <- poly.counts(flyingFox, indexData)  
indexData$Kangaroo <- poly.counts(kangaroo, indexData)  
indexData$Koala <- poly.counts(koala, indexData)  
indexData$Wombat <- poly.counts(wombat, indexData)  
  
#Add the social media data to the grid  
indexData$flickr <- unlist(extract(flickr, indexData))  
indexData$flickrResiduals <- unlist(extract(flickrResiduals, indexData))  
  
#Lastly, add the primary components   
indexData$pc1 <- unlist(extract(subset(NaturalBrickPCA, 1), indexData))  
indexData$pc2 <- unlist(extract(subset(NaturalBrickPCA, 2), indexData))  
indexData$pc3 <- unlist(extract(subset(NaturalBrickPCA, 3), indexData))  
indexData$pc4 <- unlist(extract(subset(NaturalBrickPCA, 4), indexData))  
indexData$pc5 <- unlist(extract(subset(NaturalBrickPCA, 5), indexData))  
indexData$pc6 <- unlist(extract(subset(NaturalBrickPCA, 6), indexData))

## Correlation and Simple Regression

These are both measures of relationship between continuous variables, readily identified by anyone who has taken an introductory-level statistics course. They are also very quick to generate, as they do not need to take into account the spatial nature of data.

Some things are immediately apparent:

1. To put these values to scale, even the best predictor from a PCA model held a correlation of ~0.10 at best. Clearly, this suggests that there is a real value to including social-media presence as a predictor in many of these cases.
2. The correlation of the residuals is maintained in the same direction, and only of slightly reduced power relative to the original social media variable. This suggests that the predictive power of social media presence is at least somewhat independent of the natural predictors.
3. Social media presence is more predictive of some species than others. This is harder to explain. One possibility is that the data collection methodology for some species is more rigorous than that of others.

The rest of the indices presented are focused on the Koala data, as that is the in general the one that demonstrates the strongest bias vis a vis social media presence across all the measures

## Predictive Power

This is a more applied approach to determining the predictive power of a given variable alone. Here are the outputs of a simple SDM, trained with just the flickr and re-projected flickr data.

# Distance-based (Discrete)

This metric is based off Kadmon (2004). There, the biasing predictor was whether or not a given grid cell was on the highway or not. In order to identify if there was a road-side bias, cells at a distance d from the road were compared randomly against randomly selected cells on their probability of having a recording or not. The advantage of this method over strict correlation is that the “bleeding out” of the biasing variable allows the capture of areas which may still be affected by the biasing but aren’t directly on the road itself.

This was first performed using a binarization of the variables; whether the residual social media presence was greater than its first quartile and whether there was a recording or not.

#Threshold of flickr residual score  
threshold <- unname(quantile(indexData$flickrResiduals, 0.25))  
  
#Distance from a "flickr cell" to check  
distance <- 1  
cellWithinDistanceOfRoad <- c()   
cellWithKoala <- c()  
  
  
#Load helper functions that identifies cells next to a given one  
source("Helpers/ExpandGridCell.R")  
  
#Go through each indexData cell  
for(i in 1:length(indexData)){  
 polygon <- indexData[i,]  
   
 #Track if a flickr neighbor was found  
 found <- FALSE   
   
 #Start looking in the current cell  
 searchIn <- polygon  
   
 #If the cell has a flickr score, save it as within the treshold and   
 # Skip the rest  
 if(searchIn$flickrResiduals > threshold){  
 cellWithinDistanceOfRoad <- c(cellWithinDistanceOfRoad, 1)  
 cellWithKoala <- c(cellWithKoala, polygon$Koala > 0)  
 found <- TRUE  
 } else{  
   
 #Set the maximum distance from a cell to look in  
 currentDistance <- 0  
 while(currentDistance < distance){  
   
 #Start with an empty set of new search points  
 newSearchIn <- searchIn[0,]  
   
 #For each cell in the current search space, add all surrounding cells  
 for(i in 1:nrow(searchIn)){  
 newSearchIn <- rbind(newSearchIn, searchIn[i,], expandGridCell(indexData, searchIn[i,]))  
 }  
   
 #Filter out duplicates  
 newSearchIn <- newSearchIn[!duplicated(newSearchIn$id),]  
   
 #If any of the newly added cells contain a flickr cell, stop searching.  
 #Save the cell as withn that many of the road.   
 #If the cell contains a flying fox, record that.  
 if(any(newSearchIn$flickrResiduals > threshold)){  
 cellWithinDistanceOfRoad <- c(cellWithinDistanceOfRoad, 1)  
 found <- TRUE  
 cellWithKoala <- c(cellWithKoala, polygon$Koala > 0)  
 break  
 }  
   
 #The newly identified cells are now the ones to look into  
 searchIn <- newSearchIn  
   
 #Idecrease teh distance you are  
 currentDistance <- currentDistance + 1  
 }  
 }  
   
 #If none was found after trying all poisslbe distances, quit out.  
 if(!found) cellWithinDistanceOfRoad <- c(cellWithinDistanceOfRoad, 0)  
}  
   
# Threshold of Binomial Approximation  
pD <- mean(cellWithinThresholdOfRoad)   
pF <- mean(indexData$Koala > 0)  
pR <- mean(cellWithKoala)  
N <- length(indexData)  
  
#T-Score  
(N\*pR - (N\*pF \* pD)) / sqrt(N \* (pD\*pF)\* (1-(pD\*pF)))

This methodology resulted in a t-score of of 23.15, on a distribution of 70000+ of freedom, suggesting that under this binary method, there is definitely a bias involved.

## Distance-based (Continuous)

In the continuous domain, the previous procedure is akin to performing a moving average of the social media data, prior to performing the standard correlative procedure from the first suggestion index.

#Function to generate moving averages of raster data. Should be replaced when an appropriate function in the raster package is found  
maCorrelation <- function(threshold, dataOriginal)  
{  
 dataMA <- dataOriginal  
   
  
 for(i in 1:nrow(flickrDecorrelated)){  
 for(j in 1:nrow(flickrDecorrelated)){  
   
 xIndices <- (i-threshold):(i+threshold)  
 yIndices <- (j-threshold):(j+threshold)  
   
 flickrMA[i,j] <- mean(flickrDecorrelated[xIndices, yIndices], na.rm=TRUE)  
 }  
 }  
   
 return(flickrMA)  
}  
  
#Add the moving average data to the grid  
indexData$moving1 <- unlist(extract(maCorrelation(1, flickrResiduals), indexData))  
indexData$moving5 <- unlist(extract(maCorrelation(5, flickrResiduals), indexData))  
indexData$moving10 <- unlist(extract(maCorrelation(10, flickrResiduals), indexData))  
indexData$moving25 <- unlist(extract(maCorrelation(25, flickrResiduals), indexData))  
indexData$moving50 <- unlist(extract(maCorrelation(50, flickrResiduals), indexData))  
  
#Correlation in a table  
movingAverageCorr <- data.frame(Window=c(1,5,10,25,50),  
 Correlation = cor(indexData$moving1, indexData$flickrResiduals,  
 use="complete.obs"),  
 Correlation = cor(indexData$moving5, indexData$flickrResiduals,  
 use="complete.obs"),  
 Correlation = cor(indexData$moving10, indexData$flickrResiduals,  
 use="complete.obs"),  
 Correlation = cor(indexData$moving25, indexData$flickrResiduals,  
 use="complete.obs"),  
 Correlation = cor(indexData$moving50, indexData$flickrResiduals,  
 use="complete.obs"))

## Similarity-Based

The final metric considered for this analysis was based on the statistical philosophy underlying Phillips, 2009. Rather than the traditional distance metric, this index instead considers cells to be “far apart” if their biological predictors are similar. The underlying idea being that if regions are similar biologically, we would expect them to have equivalent recording rates.

This was not completed in time for this report, as the procedures I wrote for generating distances on raster files were prohibitively expensive ((160661^2) / 2) comparisons). I also wasn’t sure what distance metrics would be appropriate. Any suggestions?

# References

Gueta, Tomer, and Yohay Carmel. “Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models.” Ecological informatics 34 (2016): 139-145.

Kadmon, Ronen, Oren Farber, and Avinoam Danin. “Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models.” Ecological Applications 14.2 (2004): 401-413.

Millar, Edward E., E. C. Hazell, and S. J. Melles. “The ’cottage effect’in citizen science? Spatial bias in aquatic monitoring programs.” International Journal of Geographical Information Science (2018): 1-21.

Phillips, Steven J., et al. “Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data.” Ecological applications 19.1 (2009): 181-197.