Methods of limiting spatial sampling bias in species distribution modeling: *Culex pipiens*, a case study

Robert Richards, Paige Miller, Gio Righi

March 25, 2016

Introduction

Ecologists, managers, and public health officials use species distribution models (SDMs) to predict the likelihood of species presence across a landscape of importance. Construction of these models requires "training" data representing known locations where the species of interest is present and absent. Presence-absence data produced through systematic samples chosen unbiasedly from the landscape of interest are the gold standard in this training data. Unfortunately examples of such data sets are few and far between. Much more common are examples of "Presence-only" data such as records from natural history museums. Though presence-only records provide less information than systematic presence-absence records (and limit the types of models that can be used) this data format can provide distinct advantages as well.

Notably, these methods are not subject to the measurement uncertainty associated with absence records. If a species is detected at a location then it is definitively present in that location at that time. If, however, a species is not detected at a given location that species may be genuinely absent, or simply have gone undetected by the researcher despite its presence. Often even true absences may bear the mark of dispersal barriers or preventive biotic interactions leading to areas that are environmentally habitable by the species being marked as absences. Though presence-only data can also bear the imprint of such non-environmental variables (species may be detected in sink habitats as a result of migration from permanently habitable source habitats) (Pulliam paper), this burden is generally considered to weigh more heavily on absence data. Presence only data are not, however, without other issues for species distribution modelers.

Generally, presence-only models rely on the assumption that presence data are sampled unbiasedly from the distribution of a species. Violation of this assumption can cause problems with both model fitting and model evaluation (Hijmans 2012). A number of methods have been proposed for addressing this problem of spatial sampling bias. They tend to fall into two major categories: (1) Thinning, a process in which a subset of presence points is selected for model fitting in such a way as to minimize spatial or environmental bias, and (2) weighting, in which points are assigned weights for model fitting based on the inverse of their association with sampling bias (e.g. Syfert et al. 2013, Stolar and Nielsen 2014). Currently this work largely focuses on the former family of methods with plans to implement novel presence-only weighting methods in the future.

Thinning has been implemented and tested in a number of different ways. The distinctions between these methods fall along 2 distinct axes: (1) the method (grid based, distance based, or cluster analysis based) used to thin the data points and (2) the axes (geographic or environmental) along which the thinning of points is performed (Fig 1). Likely the most common form of thinning performed in modern SDM is the type implemented in MaxEnt (the leading presence-background SDM). MaxEnt selects one presence point per grid-cell of the environmental covariate grid. This process can also be interpreted as MaxEnt assigning a presence or non-presence identity to each grid-cell based on whether a presence point is found inside of it. (Phillips 2006, something else too). Grid-based geographic filtering has been used with other SDMs as well (examples). Boria et al. (2014) propose and test an alternative method of geographic thinning in which a minimum allowable distance between points is imposed on presence data using an algorithm that maximizes the total number of retained points given this constraint (see Methods for further details).

Recently, efforts have been made to identify which data processing methods, including Thinning (see Table 1), provide the best model performance in the presence of spatially biased data. Some focus on the merits of geographic thining (e.g Boria et al. 2014, Kramer-Schadt et al. 2013) while others attempt to compare performance across environmental and geographic methods (Varela et al. 2014, Fourcade et al. 2014). Results of such studies have been mixed. Varela et al. (2014) compare the relative performance of grid-based environmental and spatial thinning to find the clear superiority of environmental methods. They themselves highlight, however, that this relative performance may be context specific. Their study comprised the Iberian peninsula, a particularly environmentally heterogeneous area. Environmental thinning may perform better in this setting because it retains a larger number of unique points in environmental space while geographic thinning will tend to discard neighboring points that are in fact very different environmentally. Alternatively, spatial heterogeneity of environment may result in a spatial sampling bias not manifesting as a clear bias in environmental space. This interpretation would suggest that environmental thinning ought to be less able to correct the bias in the data. Interestingly, Fourcade et al. (2014), in a comparison of a number of different methods for limiting the effect of spatial sampling bias on MaxEnt modeling, found "Systematic Sampling", a methodological equivalent of our geographic grid based thinning, to be the generally superior method. This discrepancy between results could be due to differences in the taxa modeled, the geographich region (North America for Fourcade et al.), or simply the fact that the latter group's implementation of environmental thinning was implemented using a cluster analysis, rather than grid based thinning. Though more work certainly remains to be done in this area, thinning performance is beginning to appear to be both system and implementation dependent.

Here we make an initial attempt at applying the 2 primary data thinning methods across gegraphic and environmental space at a series of resolutions in an effort to measure not the overall performance of these models but instead the similarity in the predictions that they produce. We hope to use these analyses to understand if predictions based on different data processing methods differ substantially and how they differ. We also hope to begin to tackle the problem of spatial resolution in thinning methods by quantifying both the similarity between the predictions produced by a given model at multiple resolutions and beginning to understand the ways in which these predictions tend to differ.

Methods

Data

Presence records for 3 species of Culex mosquitoes (C. pipiens, C. quinquefasciatus, and C. salinarius) on the continent of Africa were collected in the Vector Map database (http://vectormap.si.edu/dataportal.htm). Environmental covariate rasters were downloaded from the BIOCLIM database (Hijmans et al. 2005, http://www.worldclim.org/bioclim). Monthly temperature range variables were also constructed (Drake, unpublished work). Data were rescaled by substracting the raster mean and dividing by the raster standard deviation. Variables that are ecdf transformed are not rescaled.

Data Processing

Spatial Thinning

Distance-based thinning was performed using the R package spThin (Aiello-Lammens et al. 2015). The thinning algorithm accepts a minimum allowable distance (x) between points, calculates the number of occurrence records within this distance for each presence point, and identifies the record(s) with the greatest number of neighbors within x. One of these records is then eliminated at random and the process is repeated until no remaining point has a neighboer within x. This process is iterated a set number of times and the output(s) with the maximum number of remaining points are provided for training. Thinning was performed at a series of scales (x=10km, 100km, 500km) to assess the effect of thinning distance on model output.

Grid-based thinning was performed using the gridSample function in the R package dismo (Hijmans et al. 20??). We provide the function with a raster grid of a specific resolution (cell size) and (much like the MaxEnt functionality discussed above) one presence point is chosen at random in each grid cell. This type of thinning was performed at the resolution of the environmental variables (.1666 degrees) as well as lower resolutions (1 and 5 degrees). Due to the the large spatial extent of Africa the actual distance in meters corresponding to a degree varies substantially latitudinally across the continent. Nonetheless, these 3 resolutions are roughly equivalent to the three scales applied in the distance based spatial thinning.

Environmental thinning

We performed a principal components analysis (PCA) of the re-scaled baseline environmental data. Approximately 55% of the variation in environmental covariates was explained by the first two principle components (Drake, unpublished work). Thus, we sought to test whether filtering points in two-dimensional PCA space, defined by the first two principal components, is a viable option for correcting spatially biased data. Two methods of environmental thinning in PCA space were tested: (1) distance thinning, which specifies a minimum distance between any pair of points in two-dimensional PCA space and (2) grid-based thinning, which is similar to grid based thinning in geographical space, and eliminates all but one point per grid cell. We provide functions for both methods.

For distance thinning in environmental space, distances between points were calculated using the spDists function in the R package sp which constructs a vector of Euclidean distances between a matrix of 2D points. Then, we thinned data by setting a minimum distance , in PCA space, that each point had to be from any other point. Points within 1 unit from another point were eliminated from analysis. The number of points kept was maximized for each value of x.

Grid-based thinning in PCA space was performed by imposing a raster layer over the first two principal component axes and assigning a grid of specified resolution to that raster. Then, any points in the same grid cell were eliminated. A resolution of 0.1 units was used for this analysis and left around 55 points for training.

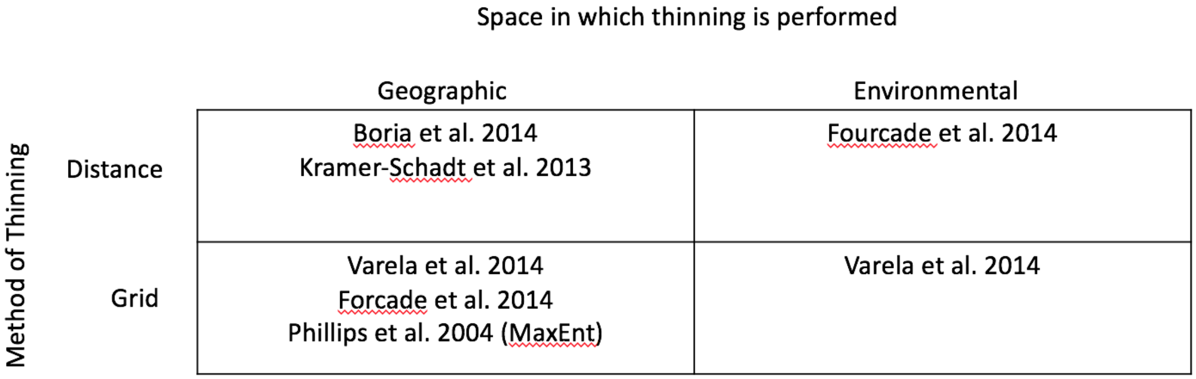
Species Distribution Models LOBAG-OC (Drake 2014) was the primary method used to model Culex distributions. Unlike the more widely used MaxEnt LOBAG-OC is a true presence-only method in that the model is trained using exclusively the point presence data, with no requirement for randomly sampled background (or "pseudo-absence") points.

Re-sampling

As a first attempt at applying a weighting approach to correcting sampling bias for LOBAG-OC models we performed biased resampling of presence points. Presence points were sampled with replacement using the sample function in base R. The likelihood of a given point being sampled was determined by characteristic weights representing the inverse of the spatial sampling bias. This underlying bias weighting can be drawn from some underlying distribution across the landscape which we suspect to cause, or be directly related to, the spatial bias in the data. In this case the points were weighted in accordance with the inverse of the population density (taken from the Socioeconomic Data and Applications Center, SEDAC) based on intuition that mosquito sampling may occur disproportionately in population dense areas. Future methods will involve tuning the resampling correction via adjusting the total number of points drawn in the sample as well as applying multiple other bias weighting factors. Finally we will look further into potential methods for generating the bias weightings from the initial distribution of presence points.

Comparison of model Predictions

Trained models were applied across the entire continent of Africa and predicted species distributions were generated for each correction method (Fig). These distributions were subsequently compared in geographic space in a pairwise fashion using Schoener's D (Schoener 1968, Warren 2008) a simple and well used ecological metric for determining the difference in probability of presence of two different species across a landscape (or series of niches). In this case we will be using the multiple different data-processing types in the place of individual species. Schoener's D ranges from 0 (no overlap) to 1 (completely identical probability predictions).



Space in which thinning is performed

Method of Thinning		Geographic	Environmental
	Distance	Boria et al. 2014 Kramer-Schadt et al. 2013	Fourcade et al. 2014
	Grid	Varela et al. 2014 Forcade et al. 2014 Phillips et al. 2004 (MaxEnt)	Varela et al. 2014

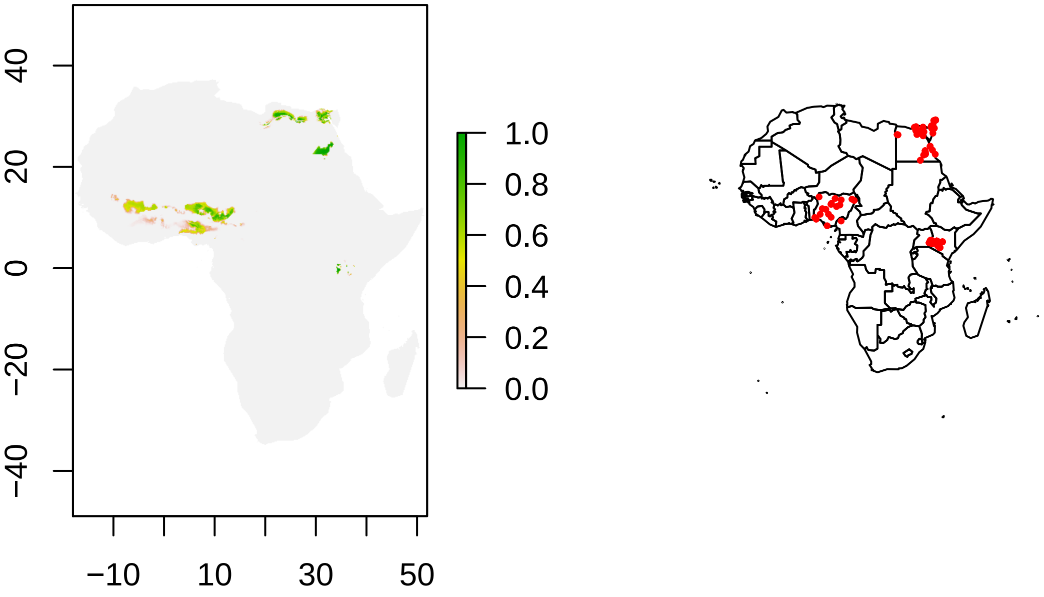
Figure 1: Table 1. Previous reports on data thinning

Results

Results

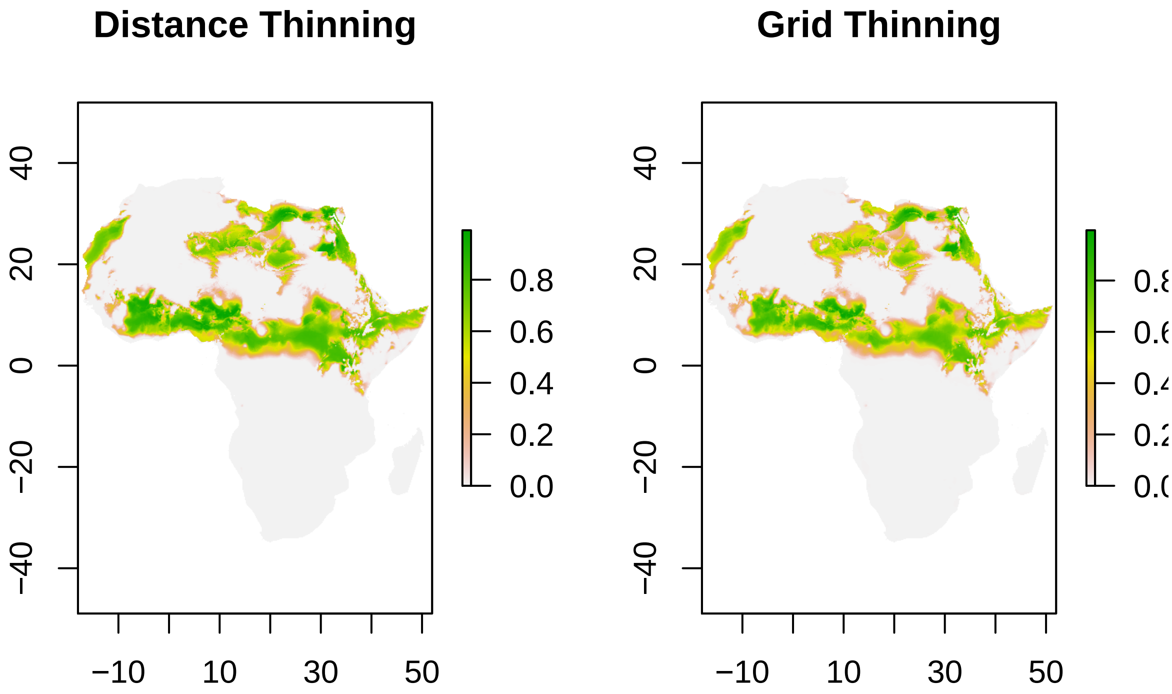
Predicted niches

Predicted distribution based on unique sampling coordinates



A LOBAG-OC model trained on all unique African presence records produces an extremely limited predicted distribution. The predicted distribution seems heavily influenced by the spatial differences in density of known occurrences. It is likely that this spatial bias is due, at least in part, to spatial bias in human sampling of mosquitoes. Therefore, as discussed above, our goal is to produce a model free from the influence of this sampling bias.

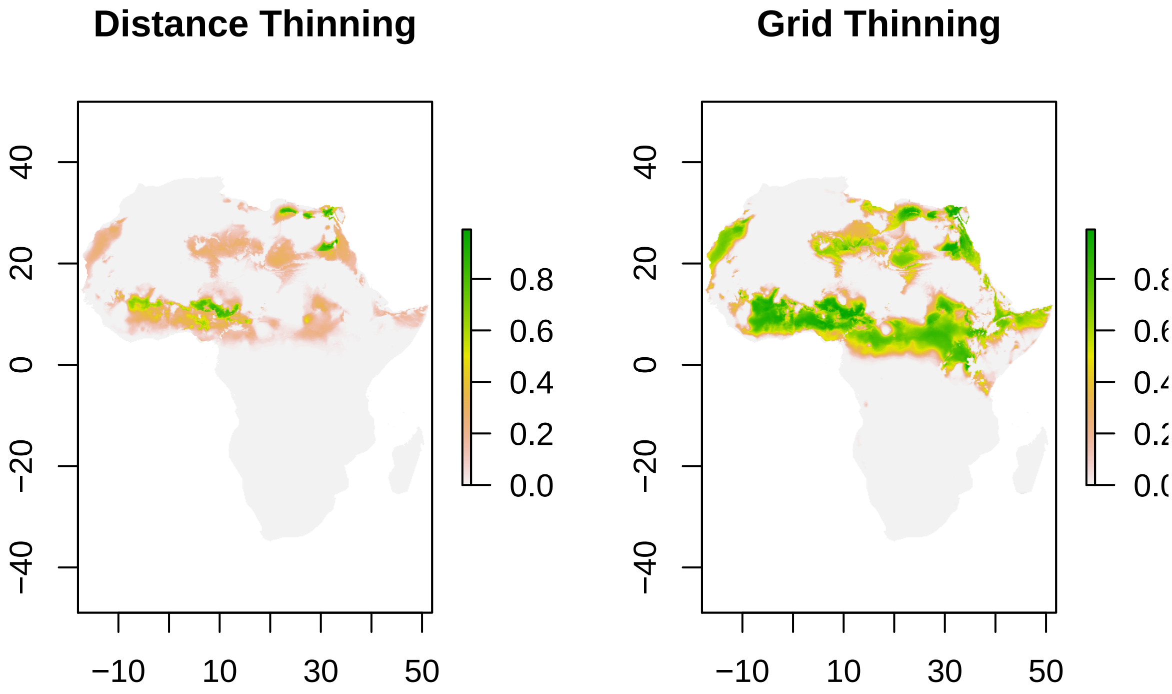
Predicted distribution for spatially filtered methods


Distance Thinning

Grid Thinning

When presence points are spatially thinned the predicted distribution expands substantially. Distance thinning at 10km and Grid thinning at the spatial resolution of the environmental data (.16 degrees) produce remarkably similar predicted distributions.

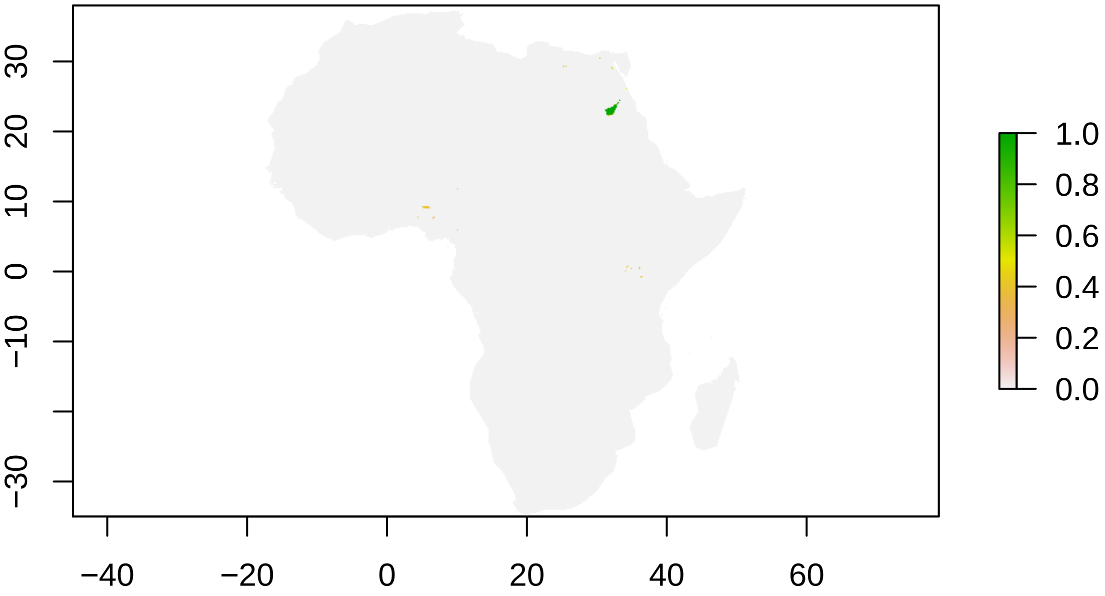
Predicted distribution for environmentally filtered methods


Distance Thinning

Grid Thinning

Our example of environmental distance thinning provides a substantially more limited predicted distribution than the grid-thinning example. Our resolution for environmental thinning is in PCA space, therefore our thinning units are somewhat arbitrary. As a result further investigation of the effects of thinning resolution on model outputs will be necessary. Particularly we hope to determine if resolution tuning can produce very similar predicted distributions from the two data-processing methods as can be seen in the geographic thinning example.

Predicted distribution method of re-sampling



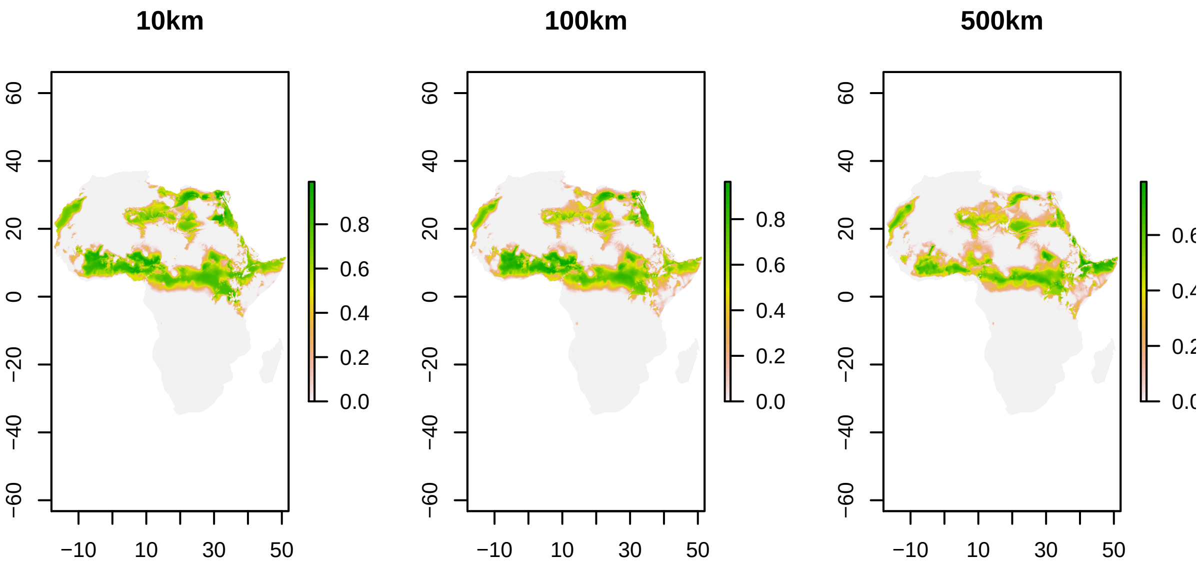
The trial counter-bias resampling method clearly did not perform as hoped. This is likely because when sampling with replacement it will be necessary to “tune” the method by setting the number of records to sample. This figure represents resampling a number of points equal to the initial number which seems to functionally increase the spatial bias of the distribution rather than reducing it.

Comparison of predicted niche overlaps

	No.Thinning	Geographic.Distance	Geographic.Grid	Environmental.Distance	Environment
No Thinning	1.000				
Geographic Distance	0.148	1			
Geographic Grid	0.154	0.969	1		
Environmental Distance	0.323	0.645	0.655	1	
Environmental Grid	0.152	0.93	0.931	0.661	1

This table reinforces the conclusions drawn above from visual inspection. All of our thinning methods are more similar to each other than they are to the no thinning case. Notably Environmental Grid thinning (at .1 resolution) appears remarkably similar to both methods of geographic thinning (10km, 0.16 degree resolution). Further work will attempt to analyze how and why these distributions differ.

Spatial distance-based thinning at multiple resolutions



10km

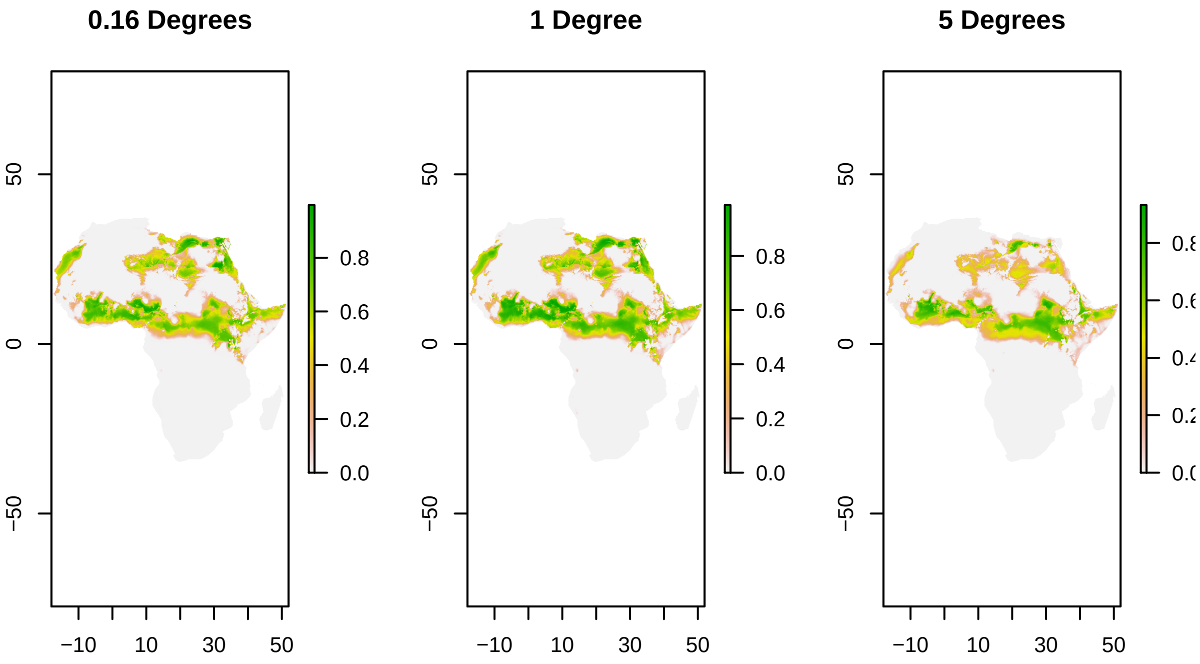
100km

500km

Schoener's D comparison of model outputs across resolutions

	10km	100km	500km
10km	1.000		
100km	0.921	1	
500km	0.783	0.797	1

Spatial grid based thinning at multiple resolutions



0.16 Degrees

1 Degree

5 Degrees

Schoener's D comparison of model outputs across resolutions

	0.16 Degrees	1 Degree	5 Degrees
0.16 Degrees	1.000		
1 Degree	0.939	1	
5 Degrees	0.803	0.819	1

These preliminary analyses of the effect of resolution on spatial thinning clearly demonstrate an effect of resolution of thining on predicted model. Under visual inspection decreased resolution (larger distances) seems to expand the total non-zero area of the predicted distribution but decrease the likelihood of presence at the heights of the distribution. Further analysis is required to better understand the relationship between resolution of thinning and model output.

Discussion

Our analysis shows that environmental and spatial filtering can produce very similar predicted distributions. This may help to address the ongoing debate about the preferred method of data thinning for Species Distribution Modeling. It seems likely that the identity of the superior method is context dependent given that in the scenario considered in the present study their results are nearly indistinguishable. For the remainder of this project we plan to further consider the effects of resolution of thinning on model output as well as further develop a weight based data-processing approach for the presence-only method LOBAG-OC. Finally we are in the process of developing a viable dataset (either simulated or real) upon which we can impose spatial bias in order to thorouhgly compare performance of our methods via AUC.

References