Correcting sampling bias in species distribution modeling: A case study of *Culex* in Africa

Robert Richards, Paige Miller, Gio Righi

May 3, 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 (e.g. J. M. Drake and Beier 2014; Hamazaki 2002; Guisan and Thuiller 2005). Construction of these models requires 'training' data representing known locations where the species of interest is present and absent (see Guisan and Zimmermann 2000 for a review). 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 or much citizen science data (C. Graham et al. 2004). 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 (Gu and Swihart 2004). 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 2000), 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 (Robert J 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 geographic or environmental aggregation, and (2) accounting for bias within the model. The second method can be accomplished by bias weighting, in which points are assigned weights for model fitting based on the inverse of their association with sampling bias (e.g. Syfert, Smith, and Coomes 2013; Stolar and Nielsen 2015). Alternatively bias can be corrected within a model by fitting the likelihood of observing a species presence as a function of both environmental variables and 'observer bias' variables (Warton and Aarts 2013).

Explicitly correcting or accounting for bias is generally considered the preferable method both because it attempts to reconstitute the unbiased distribution for modeling and because it retains all training records. Unfortunately these methods can be difficult to realize with most data sets as they tend to require a baseline understanding of the way the sampling is biased geographicly. This understanding is often derived from the bias observed in sampling distributions of other species in the same regions (e.g. Phillips et al. 2009) or basic assumptions about the way that researchers tend to sample [e.g. bias in favor of areas close to roads or population centers, (Warton and Aarts 2013; Higa et al. 2015). In the absence of such understanding data thinning provides a 'researcher behavior' free method of improving predictions in the presence of spatial sampling bias.

Thinning addresses spatial bias from a different perspective. Rather than attempting to determine the nature of the sampling bias, thinning removes aggregated points. In this way it reduces the spatial bias at a given

scale at the cost of a smaller training sample. 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 (Table 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, Anderson, and Schapire 2006) (something else). Grid-based geographic filtering has been used with other SDMs as well (Varela et al. 2014). Boria et al. (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).

		Space in which thinning is performed	
		Geographic	Environmental
Method of Thinning	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

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 (e.g. Varela et al. 2014; Fourcade et al. 2014). Results of such studies have been mixed. Varela et al. (Varela et al. 2014) compare the relative performance of grid-based environmental and geographic 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. (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 geographic region (North America for Fourcade et al. 2014), or simply the fact that the latter group's implementation of environmental thinning was a cluster analysis (choosing a single point from identified clusters, 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.

Relatively little effort has been made to assess the effect of resolution of thinning on model fit. Most thinning attempts, even those whose stated goal is to measure the effects of thinning, set a single resolution of thinning for all analyses (Boria et al. 2014; Varela et al. 2014; Fourcade et al. 2014; Phillips, Anderson,

and Schapire 2006). We instead allow the resolution of our thinning methods to vary over a range in an effort to determine the effect of thinning distance on model fit. This process allows us to compare the performance of each individual model to the 'ideal' thinning distance (the distance at which spatial autocorrelation is highest) and the model average of all thinning distances. Model averaging has been used to great effect in SDM both in the form of bootstrap aggregation (Breiman 1996; J. M. Drake and Beier 2014; J. M. Drake 2015) and of the combination of predictions of different modeling methods (Marmion et al. 2009). Therefore it is possible that if our 'ideal' thinning distance should prove inconsistent in its performance a model averaging approach will outperform an arbitrarily chosen thinning distance. Model averaging approaches often significantly outperform even the best fit models (Marmion et al. 2009).

Here we apply the 2 primary data thinning methods, geographic and environmental thinning, at a series of resolutions in an effort to measure the overall performance of each method across 30 simulated species. We then apply the best performing methods from this simulation study to 2 members of the Culex mosquito species complex in Africa whose predicted distributions have serious public health implications for the spread of West Nile Virus and whose presence data seems to be highly aggregated in geographic.

Methods

Environmental Data

Environmental covariate rasters were downloaded from the BIOCLIM database (R. Hijmans et al. 2005); (http://www.worldclim.org/bioclim). Monthly temperature range variables were also constructed (J. Drake 2013). Data were rescaled by subtracting the raster mean and dividing by the raster standard deviation. Variables that are ecdf transformed are not rescaled.

Simulated Species

The environmental distributions of N simulated species were generated using the generateRandomSp function in the R package virtualspecies using the first two principle components of our environmental covariates (Leroy et al. 2015). 2000 training presence points and 2000 testing presence and absence points were then drawn from the species' projected distribution across the continent of Africa. The distribution of training presence points was then sampled from biasedly based on proximity to 3 major population centers - Cairo (Egypt), Nairobi (Kenya), and Lagos (Nigeria) - creating a biased dataset of 500 training presence points. This process was repeated to generate 30 unique simulated species on which to test data processing methods.

Culex Data

Presence records for 2 species of Culex mosquitoes (*C. pipiens* and *C. quinquefasciatus*) on the continent of Africa were collected in the Vector Map database (http://vectormap.si.edu/dataportal.htm). Records were initially reduced to unique geographic locations, randomly sampled, biasedly sampled according to distance from the major population centers, and then were subjected to each of the following thinning methods.

Data Processing

Geographic Thinning Distance-based thinning was performed using an adaptation of 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 neighbor within x. Thinning was performed at a series of scales (x=.16 degrees, 1 degree, 5 degres) to assess the effect of thinning distance on model output. Grid-based thinning was performed using the gridSample function in the R package dismo (R J Hijmans and Elith 2015). 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 large spatial extent of Africa the actual distance in meters corresponding to a degree varies substantially latitudinally across the continent.

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 (J. Drake 2013). 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. Again two methods of thinning were tested: (1) distance thinning and (2) grid-based thinning. The implementations of these thinning methods were identical to those used for Geographic thinning.

Assessing optimal thinning resolution and Model Averaging

In order to further examine the effect of thinning resolution on model fit, we isolated environmental distance thinning as a good thinning method and performed a more thorough test of how a range of thinning distances influences model performance. We attempted, unsuccessfully, to determine an ideal thinning resolution based on linearized Ripley's K (and L) analyses. As a result we instead fit models to points that were thinned at distances in PCA space of 0.2-5 increasing at increments of 0.25. Each of the models trained on these data sets was evaluated individually and their predictions were then averaged and the performance of this consensus model was then evaluated.

Species Distribution Models

LOBAG-OC (J. M. Drake and Beier 2014) was the primary modeling method used. 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.

Model Evaluation

Models were evaluated using the 2000 presence-absence evaluation points drawn from each virtual species' distribution. Area under the Receiver Operator Curve (AUC) values were calculated and for all filtering methods were converted into delta AUC values: representing the proportion of performance lost to bias that is regained through thinning (Fourcade et al. 2014).

Culex range predictions

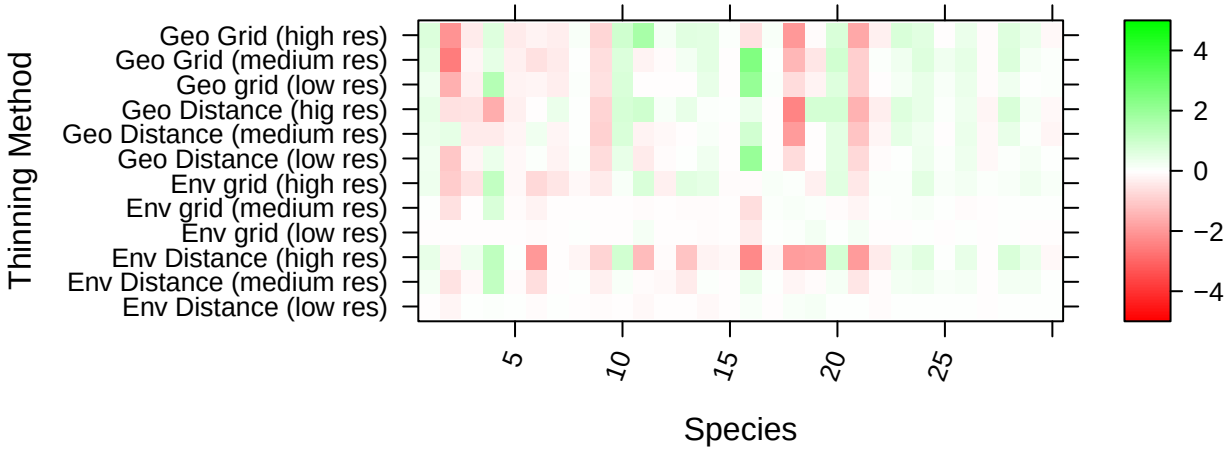
The ultimate goal of this work was to produce accurate maps of the distribution of members of the *Culex* species complex on the continent of Africa. As such we applied the mean model predictions across a range of environmental distance thinning to the geographically aggregated occurrence data mentioned above to produce predicted species distributions for *Culex pipiens* and *Culex quinquifasciatus*.

Results

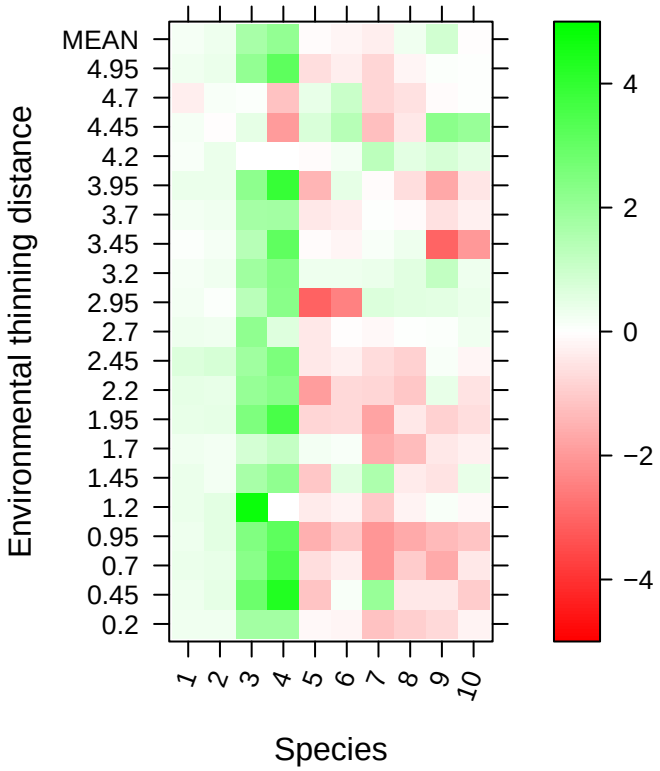
Environmental Thinning vs. Geographic Thinning

ANOVA analysis revealed the delta AUC of our 12 thinning methods (2 thinning spaces x 2 thinning methods x 3 thinning distances) significantly differed by Space (environmental or Geographic, $p<.0001$) and Distance ($p<.0001$) . Post hoc Tukey analysis reveals that all significant differences between classes were either (1)

environmental thinning methods significantly outperforming geographic thinning methods or (2) methods with a larger distance (lower resolution) outperforming methods performed in the same space with a smaller distance (higher resolution). A substantial amount of heterogeneity in model performance can nonetheless be seen across our 30 simulated species resulting in inconsistency in the best method (Figure). The best fit model for 20 species was based on data thinned in environmental space. Only 2 species had best models fit to data thinned at the highest resolution (shortest distance).



Thinning our data in environmental space at 20 distances from 0.2 to 4.95 revealed no significant pattern in delta AUC across 10 species ($p>0.05$) (Fig. 2). Instead, there was a large amount of variability across species and across each thinning distance. Some species were easier to predict (e.g. species 3 and 4), while others were harder to predict (e.g. species 5 and 8). The mean delta AUC of the averaged LOBAG models across the 20 distance thinning precedures was 0.48.



Predicted *Culex* distributions in Africa

Maps of predicted distributions for both *Culex* species show similar trends. Decreasing resolution of thinning tends to increase the extent of predicted distributions.

```
#chunk
```

Discussion

This work highlights the relative superiority of data thinning in environmental space for ameliorating the effect of spatial sampling bias. We therefore agree with the results of (Varela et al. 2014) and extend them from simple grid based thinning to distance based thinning as well. Our conclusions differ from those of (Fourcade et al. 2014) who found 'systematic sampling,' the equivalent of our geographic thinning, to be the superior method. This discrepancy likely stems from the use of cluster analysis for environmental thinning in that work. Clusters of points in environmental space were identified and then a single point was selected from each cluster. This may have led to the environmental thinning of (Fourcade et al. 2014) discarding far more points than our grid and distance based methods resulting in lower AUC values. This superiority of environmental thinning may be due to the fact that geographic thinning has the potential to discard points that, despite being close geographically, are quite environmentally distinct and therefore each offer substantial information (Varela et al. 2014). This may be particularly likely across a highly heterogeneous landscape. Alternatively, because our 3 chosen resolutions were relatively arbitrary, we may have simply chosen more ideal resolutions for environmental thinning than for geographic. Additionally the best fit model for 10 out of 30 simulated species was, nonetheless, based on geographically thinned data. Therefore we tentatively recommend thinning spatially biased data in environmental space via either grid or distance methods. Alternatively, analyses could thin in both environmental and geographic space and compare model predictions.

When we examine environmental distance thinning at a broader range of resolutions we no longer observe any significant relationship between thinning distance and DAUC. In addition there is no clear optimal environmental thinning resolution across all 10 simulated species. Therefore, we employed a consensus based method of model averaging in order to draw on the strengths of the models built at each resolution. Due to the relatively strong performance of our consensus model we recommend model averaging across a viable of range of thinning distances when possible. This method is particularly important for environmental thinning because there are no, clear, biologically meaningful units of distance in pca space with which to choose a single thinning distance. We expect this rationale to apply to geographic thinning as well. Despite the ability to select a thining distance in geographic space that is likely to be biologically relevant (e.g. dispersal distance of a focal species) the risk of choosing poorly can be mitigated through consensus methods.

We applied environmental distance thinning, as a putative best method, to data for *Culex pipiens* and *Culex quinquefasciatus* in the same manner as our in depth analysis of this method. These two species had highly aggregated presence points around Cairo (Egypt), Lagos (Nigeria), and Nairobi (Kenya). Without pre-processing, model predictions were unrealistically narrow. After thinng environmentally at a range of distances before model training, LOBAG produced vastly different predictions. This concurs with our simulation study which examined at a range of 20 distances but found no clear optimum. Thus, we present the prediction of our consensus model as our most reliable prediction of present *Culex* distributions.

Warton et al. (2013) present another promising idea for correcting spatial bias in species distribution modeling using a model-based appraoch in which the exact observer bias is not necessary. Instead they let presence locations be a point-process function of known observer bias variables (e.g. site accessibility) in addition to environmental variables. Although this method should be validated with data from more than one species, we believe this is an appealing solution to the issue of sampling bias in SDM.

References

Aiello Lammens, Matthew E, Robert A Boria, Aleksandar Radosavljevic, Bruno Vilela, and Robert P Anderson. 2015. "spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models." *Ecography* 38 (5): 541–45.

Boria, Robert A, Link E Olson, Steven M Goodman, and Robert P Anderson. 2014. "Spatial filtering to reduce sampling bias can improve the performance of ecological niche models." *Ecological Modelling* 275 (March): 73–77.

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 24 (2): 123–40.

Drake, John. 2013. "Unpublished work."

Drake, John M. 2015. "Range bagging: a new method for ecological niche modelling from presence-only data." *Journal of The Royal Society Interface* 12 (107): 20150086.

Drake, John M, and John C Beier. 2014. "Ecological niche and potential distribution of Anopheles arabiensis in Africa in 2050." *Malaria Journal* 13 (1): 1–12.

Fourcade, Yoan, Jan O Engler, Dennis Rödder, and Jean Secondi. 2014. "Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias." *PLoS ONE* 9 (5): e97122.

Graham, C, S Ferrier, F Huettman, C Moritz, and A PETERSON. 2004. "New developments in museum-based informatics and applications in biodiversity analysis." *Trends in Ecology & Evolution* 19 (9): 497–503.

Gu, W D, and R K Swihart. 2004. "Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models." *Biological Conservation* 116 (2): 195–203.

Guisan, Antoine, and Wilfried Thuiller. 2005. "Predicting species distribution: offering more than simple habitat models." *Ecology Letters* 8 (9): 993–1009.

Guisan, Antoine, and Niklaus E Zimmermann. 2000. "Predictive habitat distribution models in ecology." *Ecological Modelling* 135 (2-3): 147–86.

Hamazaki, Toshihide. 2002. "SPATIOTEMPORAL PREDICTION MODELS OF CETACEAN HABITATS IN THE MID-WESTERN NORTH ATLANTIC OCEAN (FROM CAPE HATTERAS, NORTH CAROLINA, U.S.A. TO NOVA SCOTIA, CANADA)." *Marine Mammal Science* 18 (4): 920–39.

Higa, Motoki, Yuichi Yamaura, Itsuro Koizumi, Yuki Yabuhara, Masayuki Senzaki, and Satoru Ono. 2015. "Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort." *Diversity and Distributions* 21 (1): 46–54.

Hijmans, R J, and J Elith. 2015. "Species distribution modeling with R."

Hijmans, R, S Cameron, J Parra, P Jones, and A Jarvis. 2005. *WORLDCLIMa set of global climate layers (climate grids).* International Journal of ….

Hijmans, Robert J. 2012. "Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model." *Ecology* 93 (3): 679–88.

Kramer Schadt, Stephanie, Jürgen Niedballa, John D Pilgrim, Boris Schröder, Jana Lindenborn, Vanessa Reinfelder, Milena Stillfried, et al. 2013. "The importance of correcting for sampling bias in MaxEnt species distribution models." *Diversity and Distributions* 19 (11): 1366–79.

Leroy, Boris, Christine N Meynard, Céline Bellard, and Franck Courchamp. 2015. "virtualspecies, an R package to generate virtual species distributions." *Ecography*, June, n/a–/a.

Marmion, Mathieu, Miska Luoto, Risto K Heikkinen, and Wilfried Thuiller. 2009. "The performance of state-of-the-art modelling techniques depends on geographical distribution of species." *Ecological Modelling* 220 (24): 3512–20.

Phillips, Steven J, Robert P Anderson, and Robert E Schapire. 2006. "Maximum entropy modeling of species geographic distributions." *Ecological Modelling* 190 (3-4): 231–59.

Phillips, Steven J, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. 2009. "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data." *Ecological Applications* 19 (1): 181–97.

Pulliam, H R. 2000. "On the relationship between niche and distribution." *Ecology Letters* 3 (4): 349–61.

Stolar, Jessica, and Scott E Nielsen. 2015. "Accounting for spatially biased sampling effort in presence-only species distribution modelling." *Diversity and Distributions* 21 (5): 595–608.

Syfert, Mindy M, Matthew J Smith, and David A Coomes. 2013. "The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models." *PLoS ONE* 8 (2): e55158.

Varela, Sara, Robert P Anderson, Raul Garcia-Valdes, and Federico Fernandez-Gonzalez. 2014. "Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models." *Ecography* 37 (11): 1084–91.

Warton, David, and Geert Aarts. 2013. "Advancing our thinking in presence-only and used-available analysis." *The Journal of Animal Ecology* 82 (6): 1125–34.