A Site is Not a Centroid

Modeling Archaeological Landforms with Kernel Logistic Regression on Focal Mean Embeddings

Matthew D. Harris, ***AECOM***

http://www.matthewdharris.com

@md_harris

KLRfome
https://github.com/mrecos/klrfome

Code and non-sensitive data used for this study are openly available online at:

https://doi.org/10.5281/zenodo.1211182

mrecos/klrfome

*Special thanks to **Zoltán Szabó** (Center of Applied Mathematics, École Polytechnique, Paris, France) for inspiring this approach and kindly enduring a long sporadic email dialog.

Thanks to **Chester Cunanan** (AECOM) for poster design and support.

All errors, misrepresentations, and omissions are solely my own.

1 PROBLEM: MODELING THE RICHNESS OF LANDSCAPES

PROBLEM STATEMENT

Landscape level archaeological sensitivity models are typically conceptualized as projecting the landform patterns observed at known sites to unsurveyed areas. However, most methods fail to characterize the richness of these landforms and the environmental variation is typically lost. Most commonly variation is lost when the entire landscape within a site is reduced to a single point at the center or summarized as to a mean for each variable measured. Alternatively, the variation is retained as independent observations per site area thereby violating the assumption of identically and independently distributed (i.i.d) observations due to spatial correlation.

While either of these two approaches can produce successful models, they beg the question: *Can we model a representation of a landforms Richness without losing variation?* More technically, the question is can we approximate a function to map distributions to presence/absence without assuming the shape of the distribution or costly density estimation? The answer presented here is to incorporate the methods of Distribution Regression (*Szabó et al. 2016*) along with focal Kernel Logistic Regression and focal window prediction.

TWO STAGE SAMPLING

A two stage sampling strategy is a key to modeling the distribution of a feature (e.g. distance to water) as representing a single label; this is the *Distribution Regression* problem (*Szabó et al. 2016*). This approach assumes that all site locations are samples from some meta-distribution `M` and that each site is represented by a distribution of features and a label (e.g. site presence or absence). Since we do not measure the distribution directly, we take samples to approximate it.

Thus, we take representative samples from the distribution on a site or background that are themselves samples from the meta-distribution of all possible sites and background (see Equation 1). Aggregating the intra-site observations to the inter-site level mitigates the issue of intra-site spatial correlation and moves the i.i.d assumption to the site level. This is the two-stage sampling approach. From this, the representative samples from within sites and background can be used to model the similarity between all sites and background.

MODELING FEATURE SPACE

The principle concept to this approach is comparing landforms in their feature space as opposed to geographic space; this is called the "**Kernel Trick**" in Machine Learning. Once mathematically projected into feature space, higher dimensional nonlinear functions can be approximated without explicit (and expensive) mapping; that is the trick!

In high dimensional space, a linear model is able separate complex data. Once projected back to low dimensional space, the separation is non-linear. This property makes kernel methods convenient and computationally tractable.

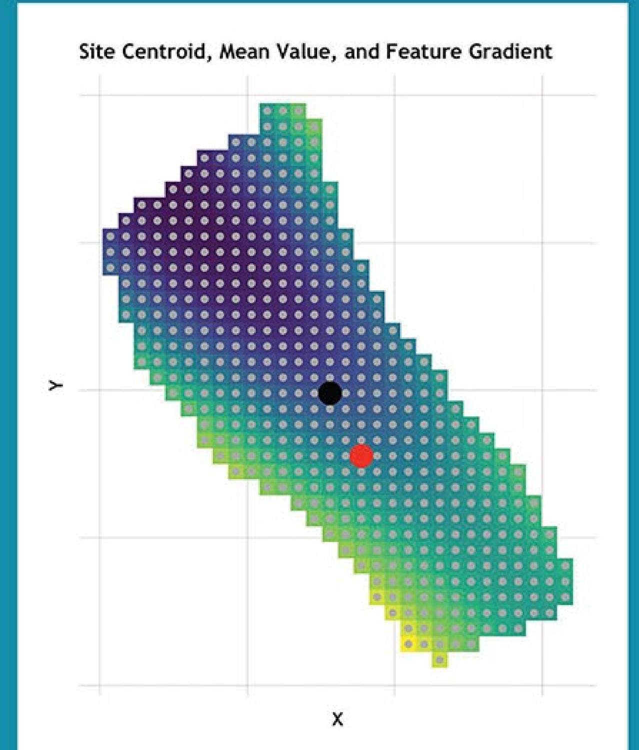


FIG. 1 - *Example of an archaeological site centroid (black dot), location of mean value (red dot), and potential sampling locations (gray dots) over an environmental gradient*

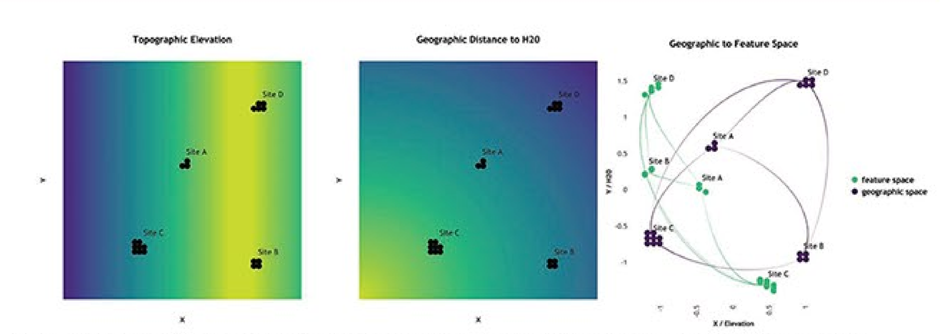


FIG. 2 - *Bivariate transformation of site locations from geographic space into feature space. Note change in relationships between sites.*

EQUATION 1
TWO-STAGE SAMPLING AND THE KERNEL TRICK

data; presence/abstract; transformed mean; distance function; Kernel; similarity function

$$\left(\left\{x_i^j\right\}_{j=1}^{N_1}, y_1\right), \left(\left\{x_i^j\right\}_{j=1}^{N_2}, y_2\right), \ldots, \left(\left\{x_i^j\right\}_{j=1}^{N_n}, y_n\right); \; y \in \{0,1\}$$

$$\hat{\mu}_{ij} = \frac{1}{N}\sum_{i,j}^{N} \phi\left(x_i^j, x_i^{j'}\right)$$

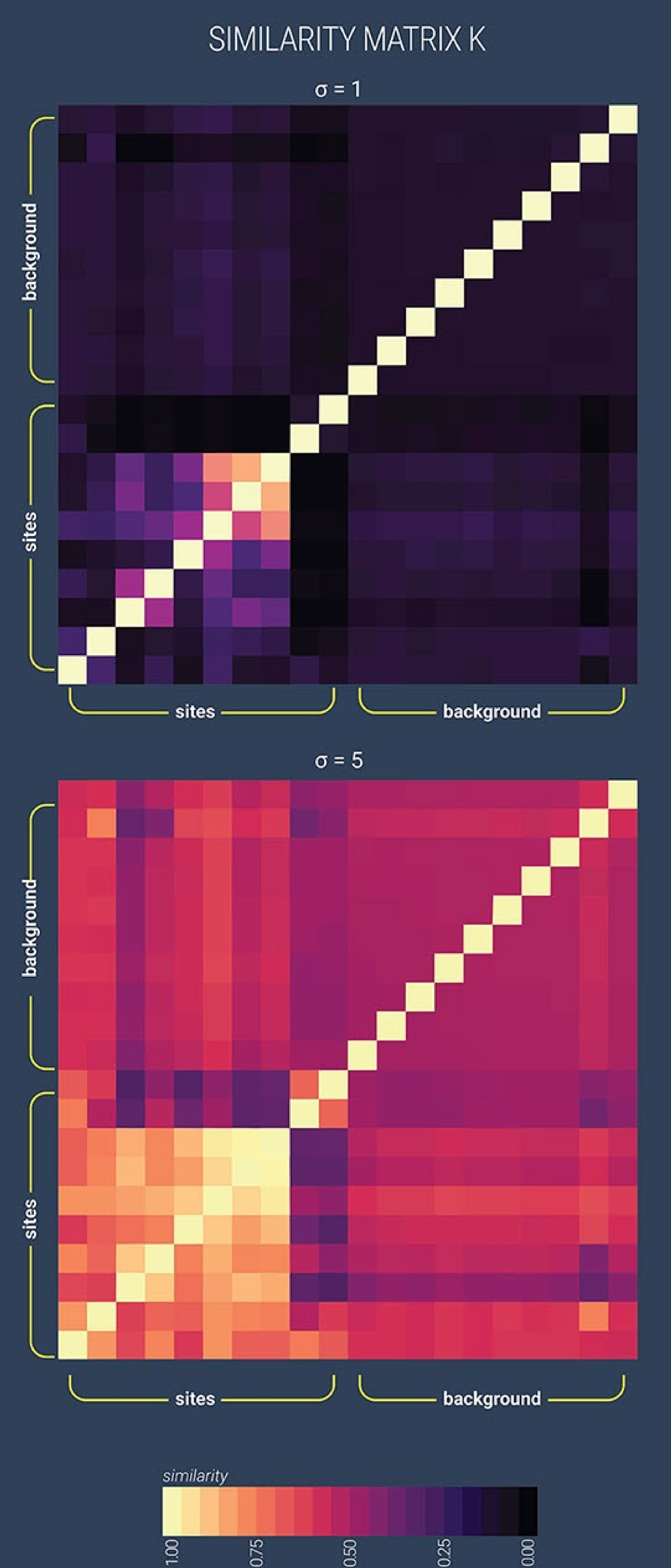
$$\phi(x_i^j, x_i^{j'}) = e^{\frac{-C(x_i^j, x_i^{j'})^2}{2\sigma^2}}$$

$$C(x_i^j, x_i^{j'}) = \sqrt{\sum_{k=1}^{ncol} (a\,(i,k) - b\,(i',k))^2}$$

$$K = K\left[\hat{\mu}_{ij}, \hat{\mu}_{ij'}\right]_{NXN}$$

$$y = f(K) + \epsilon$$

2 KERNEL LOGISTIC REGRESSION (KLR) ON FEATURE SPACE MEAN EMBEDDING



SIMILARITY KERNEL

Unlike most statistical learning approaches, this approach does not model the individual observations of presence and absence. Instead, via the two-stage sampling approach and kernel trick, it models the similarity between all sites and the background environment. This similarity is based on many samples of the characteristics that define the landscape of each site and the general environment in which they are found.

By modeling similarity, we are explicitly trying to project the same rich landscapes on which we have found sites to new areas of similar richness? Depending on the research question, specific definitions of "similarity" or "distance" can be used to model the relationship between sites. This study uses the Exponentiated Quadratic (a.k.a Gaussian kernel) to define similarity. With the use of a similarity matrix as the objective, as opposed to simple tabular data, a Kernel Logistic Regression (KLR) approach (*Zhu and Hastie 2005*) is used.

LEARNING PARAMETERS WITH NEWTON'S METHOD (IRLS)

In order to project the similarity of known sites into areas that have not been surveyed we must learn a set of functions to fit our training data in the similarity kernel. Statistical and Machine Learning are terms for the process of learning functions from existing data in order to apply them to new data. In this research, we use the Iterative Reweighted Least Squares (IRLS) (aka Newton's Method) to fit the model parameters.

Generally, the IRLS approach iteratively minimizes the difference between old and new parameters regularized by the lambda hyperaparemeter. The result of IRLS is a set of parameters that can be used to compare the similarity of any new location to those that we already know and produce a probability score of whether is should be in the site-present or background class.

PROJECTION OF SIMILARITY ACROSS THE LANDSCAPE

Prediction involves the calculation of a new similarity kernel between the new test data and all of the training data. The new prediction kernel is then multiplied by the learned parameters to derive a probability of site-presence. In this research, prediction is carried out as a roving focal window function across the study area.

This means instead of predicting based on the information contained in single raster cell, the prediction is based on a N by N neighborhood around each cell; in this case 3 by 3 cells. In this sense, the prediction is that of a landscapes similarity and the view of this landscape is controlled by the size of the N by N window. The larger the focal window, the greater the perspective of the landscape and the smoother the prediction.

EQUATION 2
ITERATIVELY REWEIGHTED LEAST SQUARES (IRLS)

working response; Kernel; alpha parameters; probability function; estimated response

$$p = 1/\left(1 + e^{-K\alpha}\right)$$

$$\alpha = [1/N]$$

$$z = K\alpha + (y - p)/W$$

$$W = diag\ (p(1 - p)))_{[N]}$$

$$\alpha^{new} = (K + \lambda I)^{-1} z$$

$$\hat{f}(K) = \arg\min \sum_{i} \left(\alpha^{new} - \alpha\right) + \lambda \|f\|_{H_K}^2$$

$$\hat{y} = 1/\left(1 + e^{-K\alpha^{new}}\right)$$

EQUATION 3
FOCAL FUNCTION PREDICTION ACROSS THE LANDSCAPE

Kernel; predicted probability; alpha parameters; transformed mean

$$\hat{y}_* = 1/\left(1 + e^{-k_*\alpha}\right)$$

$$k_* = K(\hat{\mu}_*, \hat{\mu})$$

3 FOCAL PREDICTION

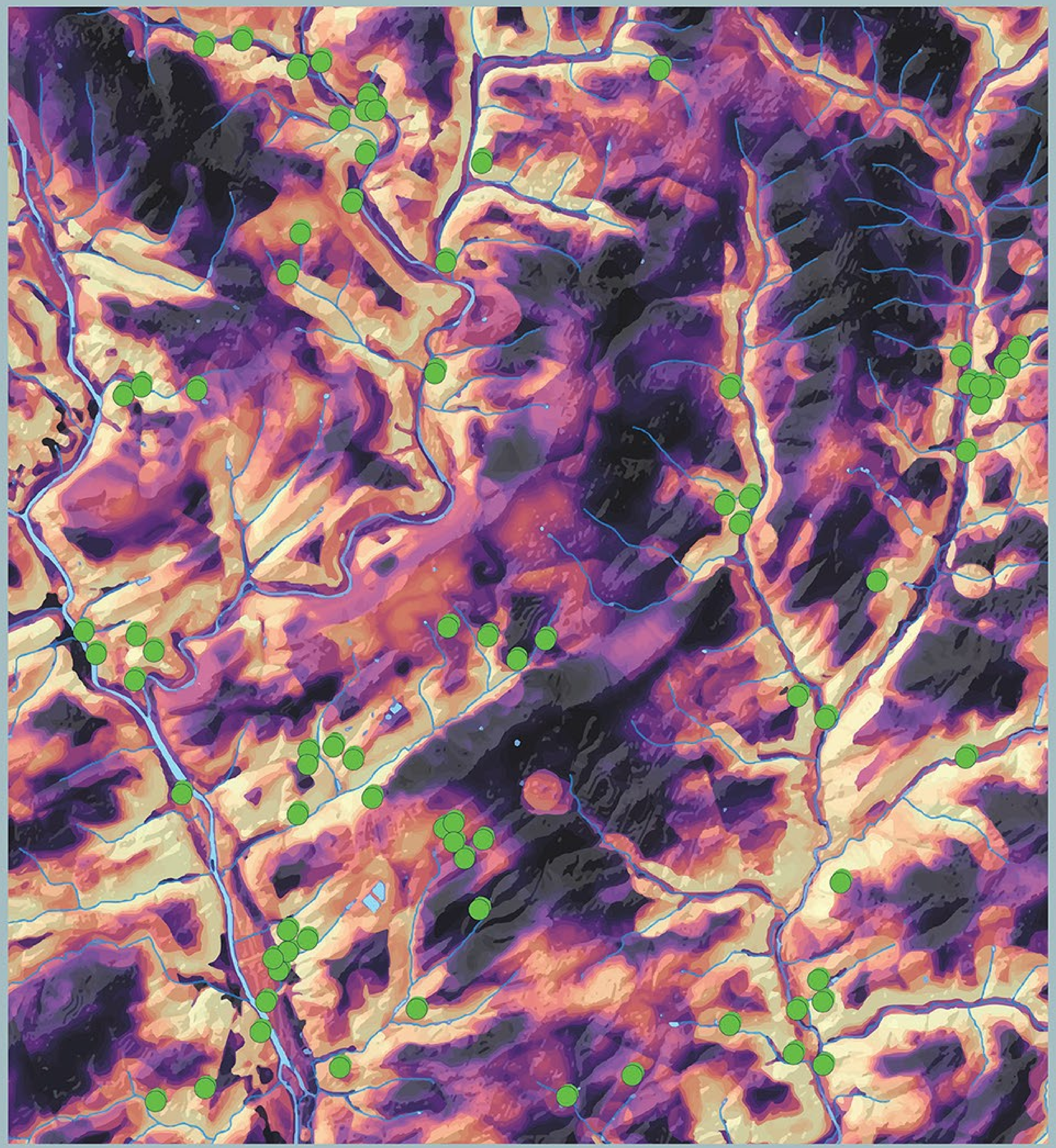


FIG. 4 - *Projection of KLRfome model onto geographic space using roving focal window with 3 by 3-cell*

4 MODEL VALIDATION AND COMPARISON

COMPARING KLRFOME TO OTHER MODELS

It is important to know how well this approach works when compared to other common or state-of-the-art models. Figure 5 is a comparison between the results of the KLR model against both Logistic Regression (LR) and a Support Vector Machine (SVM) with Gaussian Kernel. These models are compared on two different metrics (Area Under the Curve (AUC) and Youden's J), within five different physiographic zones, and over 100 different runs of the models on random subsets of the data. LR is used as a trustworthy and commonly known benchmark, while the SVM is used because it is the most well-known machine learning algorithm that also uses the kernel-trick.

In this comparison, KLR did equally as well based on the AUC and Youden's J metrics suggesting that this approach does not suffer in accuracy versus well-known models (Table 1). Interestingly, the KLR model does appreciably better than better than LR and SVM in physiographic region 1, which is the most difficult to model.

TABLE 1
Median Area Under the Curve Metric for Model Comparison

model	1	2	6	8	12
KLR	0.707	0.815	0.766	0.819	0.866
LR	0.618	0.867	0.740	0.834	0.892
SVM	0.609	0.850	0.745	0.811	0.868

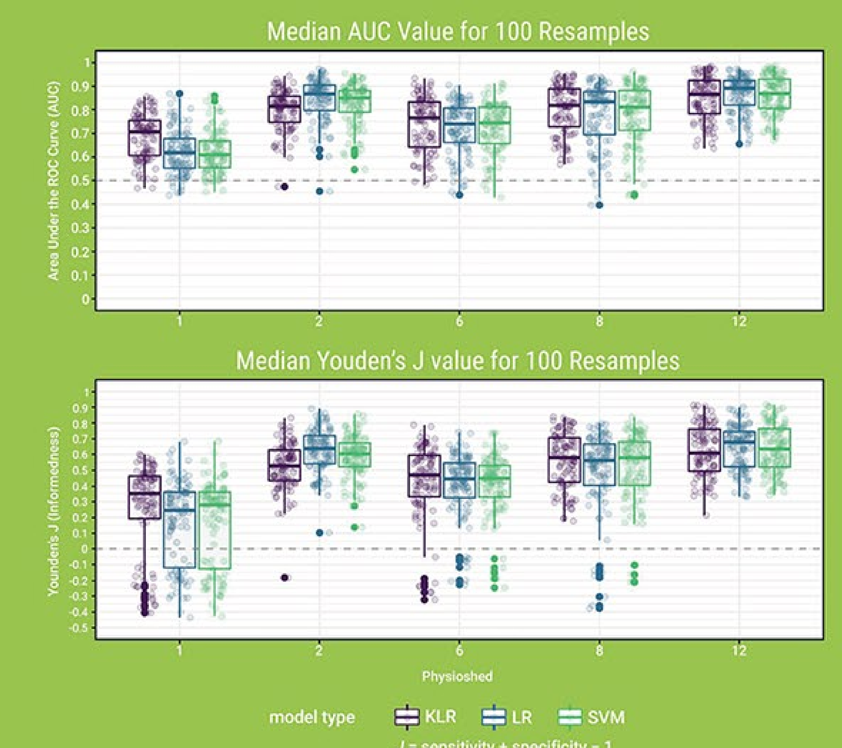


FIG. 5 - *Comparison of AUC and Youdens'J performance metrics for KLR, Logistic Regression (LR), and Support Vector Machine (SVM) models in different geographic settings.*

J = sensitivity + specificity − 1

TESTING KLRFOME MODEL PERFORMANCE

Before the model is projected onto the study area, there are numerous internal tests of model performance and validity based on held-out test samples and the 100 random resamples shown in Figure 5. These tests give a very good approximation on how the model performs in a general sense. However, in use, most sensitivity models are thresholded into high, moderate, and low sensitivity and the true performance of a model depends not only on these sensitivity classes, but also on a utility function defining "cost" for errors. When defining appropriate threshold, it is very informative to define the model across many possible thresholds and minimize or maximize the desired utility.

Table 2 illustrates the true/false positive and true/false negative quantities, as well as relevant metrics for each potential probability threshold. If maximizing for the Youden's J statistic, the optimal threshold between site-present and site-absent classes is around 0.4. At this threshold, the KLR model given this area and parameters correctly classified 91% of the site-present cells into an area covering 53% of the study area. In practice, a finer sequence of thresholds would be calculated, and a weighted utility function would be the objective.

TABLE 2
Threshold Based Performance Metrics of Geographic Prediction

Threshold	TP	FP	TN	FN	AUC	YoudensJ	KG	Sensitivity	FPR
0	9380	5000	0	0	0.762	0.000	0.000	1.000	1.000
0.1	9035	4094	906	345	0.762	0.144	0.150	0.963	0.819
0.2	8894	3529	1471	486	0.762	0.242	0.256	0.948	0.706
0.3	8799	3048	1952	581	0.762	0.328	0.350	0.938	0.610
0.4	8515	2631	2369	865	0.762	0.382	0.420	0.908	0.526
0.5	7674	2205	2795	1706	0.762	0.377	0.461	0.818	0.441
0.6	6899	1798	3202	2481	0.762	0.376	0.511	0.736	0.360
0.7	6258	1432	3568	3122	0.762	0.381	0.571	0.667	0.286
0.8	5189	1015	3985	4191	0.762	0.350	0.633	0.553	0.203
0.9	3406	505	4495	5974	0.762	0.262	0.722	0.363	0.101
1	0	0	5000	9380	0.762	0.000	NaN	0.000	0.000

5 BAYESIAN INTERPRETATION OF KLR

BAYESIAN INTERPRETATION TO QUANTIFY UNCERTAINTY

Beyond creating and characterizing a model of explicit similarity modeling, an objective of this study is to recast the model into a Bayesian framework to better propagate uncertainty into the posterior predictive distribution. In this approach, uncertainty in model parameters, location measurements, and sample sizes can be approximated with distributions on order to better understand the implications of what we *do not* know.

Presented here is the initial framework of a Bayesian KLR and an illustration of the predictive distribution. Coded in the Stan probabilistic Programming Language, this model places a range of many possible values over the lambda, alpha, and sigma parameters from the KLR model. The Bayesian model is still a work in progress.

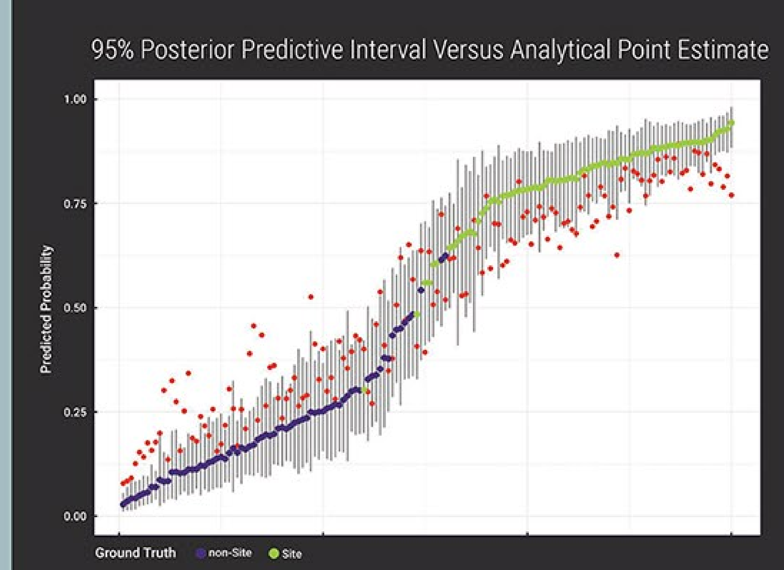


FIG. 6 - *Bayesian model 95% Credible Intervals and mean for test observations*

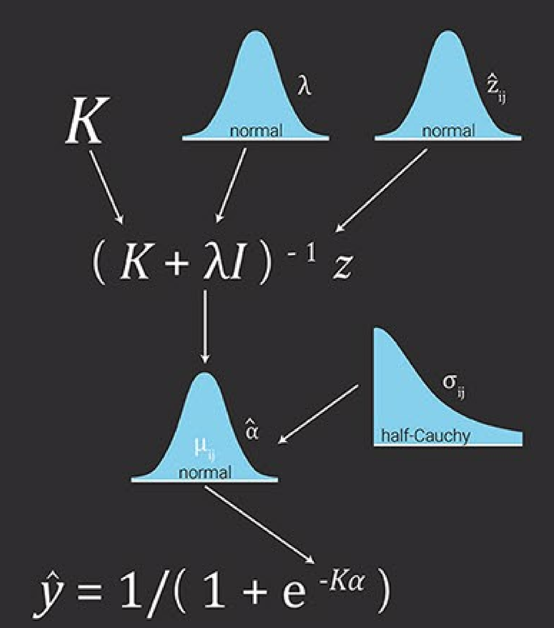


FIG. 7 - *Schematic of probabilistic model for estimating uncertainty in KLR model.*

REFERENCES

Flaxman, Seth, Yu-Xiang Wang, and Alexander J Smola. Who Supported Obama in 2012?: Ecological inference through distribution regression. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 289–298. ACM, 2015.

Law, BHo Chung Leon, Dougal J. Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian Approaches to Distribution Regression, 2017. https://arxiv.org/abs/1705.04293 [accessed 03/21/2018]

Sciaini, Marco, Matthias Fritsch, Craig E. Simpkins (2018). {NLMR}: Simulating neutral landscape models. R package version 0.2.0. URL https://CRAN.R-project.org/package=NLMR

Song, Li, Kenji Fukumizu and Arthur Gretton. Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models, in *IEEE Signal Processing Magazine*, 30(4):98-111, 2013.

Szabó, Zoltán, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In International Conference on Artificial Intelligence and Statistics (AISTATS), pages 948–957, San Diego, California, USA, 9-12 May 2015.

Szabó, Zoltán, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. Journal of Machine Learning Research, 17(152):1–40, 2016.

Zhu, Ji and Trevor Hastie. Kernel logistic regression and the import-vector machine. Journal of Computational and Graphical Statistics, 14(1):185–205, 2005.