Discovering Different Regimes of Biodiversity Support Using Decision Tree Learning

T. F. Stepinski[1], D. White[2], J. Salazar[3]

[1]Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA
Telephone: +1 513 .556.3583
Fax: +1 513.556.3370
Email: stepintz@uc.edu

[2]US Environmental Protection Agency, Corvallis, OR 97333, USA
Email: whitede@onid.orst.edu

[3]Lunar and Planetary Institute, Houston, TX 77058, USA
Email: salazar@lpi.usra.edu

1. Introduction

A pressing problem in biodiversity studies is to find the optimal strategy for protecting the species given limited resources. In order to design such a strategy it is necessary to understand associations between spatial distribution of biodiversity and environmental factors. A relationship between a response variable (a suitable measure of biodiversity) and predictor variables (measures of environmental factors) is certain to be complex as it must reflect a non-stationary character of an observed dependence. As a result one can expect an existence of several different biodiversity regimes – sets of environmental conditions *locally* associated with the levels of biodiversity measure. Multi-regime association cannot be discovered using standard methods based on linear regression; here we propose using decision tree learning methodology to discover different regimes of association between environmental variables and richness of bird species (a particular measure of biodiversity) across the contiguous United States.

Fig.1 shows a map depicting spatial distribution of richness (R) of bird species across the US. Distribution of R has a strong bimodal character effectively dividing the United States into high richness (*HR*) and low richness (*LR*) regions using a threshold value of R=148; this value corresponds to a location of the minimum that clearly separates the two maxima of bimodal distribution of R. The *HR* region is not simple-connected; instead it consists of several geographically distributed pieces. The premise is that observed distribution of R associates with locally-specific combination of values of environmental variables. We find those associations using a data mining technique based on decision tree learning. This is an expansion of a method proposed by White and Sifneos (2002).

2. Methods

We consider a set of 32 predictor variables pertaining to terrain, climate, landscape metrics, land cover, and environmental stress and hypothesized to have potential influence on bird richness. These variables constitute a subset of a larger dataset (White et al., 1999) and are given on a grid consisting of 12,337 hexagons covering the contiguous United States. A value of response variable R is the count of unique species in every hex. Breeding Bird Survey (BBS) grids (Sauer et al. 1995) representing distribution of individual bird species was used to calculate R; the values range from R=21 to R=230.

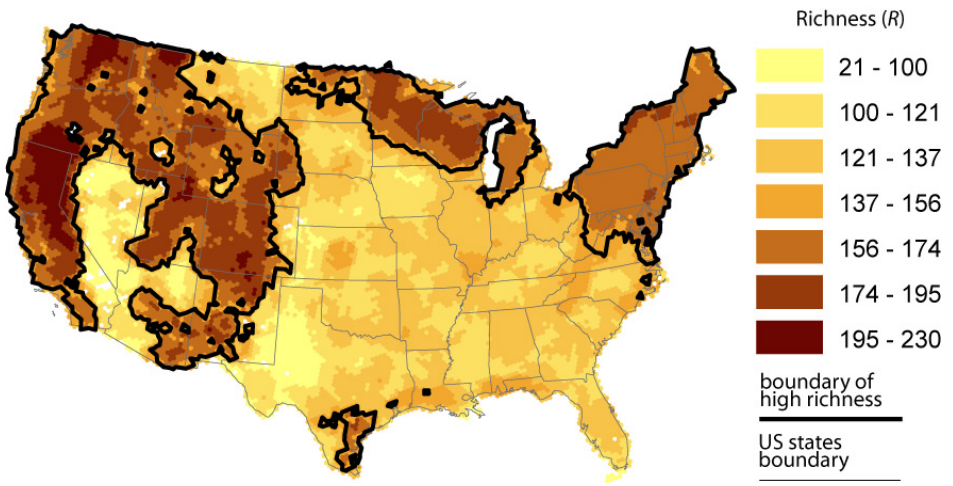

Figure 1. Map of richness, *R*, of bird species across the contiguous United States.

Data mining technique of decision tree learning (Loh, 2008) uses a decision tree as a predictive model. The model recursively partitions a set of predictor variables until each partition, represented as a terminal node of the tree, contains only data instances (hexes in our dataset) from which a conclusion about the response variable can be made with relatively high accuracy. A unique feature of the tree model is its interpretability; other models often possess good prediction accuracy, but they act like black boxes and do not provide insight into the roles of the predictor variables. Our focus here is *not* on a predictive accuracy of such model (after all, the values of *R* are known for every hex) but rather on the data partitions that we connect with different biodiversity regimes. We build two conceptually different models. First, we build a *regression tree* model which is a piecewise constant estimate of a regression function. Data is partitioned so as to increase the accuracy of linear regression in each partition. In each terminal node an average value of *R* serves as a predictor. Nodes are labeled as *HR* if they contain predominantly high values of *R* and *LR* if they contain predominantly low values of *R*. Second, taking advantage of a bimodal character of the distribution of *R*, we start by labeling hexes into *HR* and *LR* using a threshold value of R_{thres}=148, and then build a *classification tree*. In classification tree data is partitioned so as to increase the label purity of subdivisions. Nodes are labeled as *HR* if they contain majority of *HR* hexes and *LR* if they contain majority of *LR* hexes. We used GUIDE algorithm (Loh, 2008) to build regression and classification trees having 12 terminal nodes each. The number of terminal nodes is determined automatically by a process of cross validation.

3. Results

Results of the regression tree model (*RTM*) are shown on Fig.2. *HR* nodes and spatial regions corresponding to them are shown in warm colours while *LR* nodes and spatial regions corresponding to them are shown in cool colours. The overall accuracy of the *RTM* is ~80%. The major split of dataset is on the value of July mean temperature. Hexes with July temperatures <= 21.8 C are conducive to *HR*; all but one node in the left main fork of the tree are *HR* nodes and there are no *HR* nodes in the right main fork of the tree. Surprisingly, despite a complex character of the dataset, great majority of "higher richness" hexes fulfil a single (JulyMeanTemp <= 21.82) predicate. Each *HR* node groups predominantly *HR* hexes and thus can be identified with a particular environmental regime conducive to high richness of species.

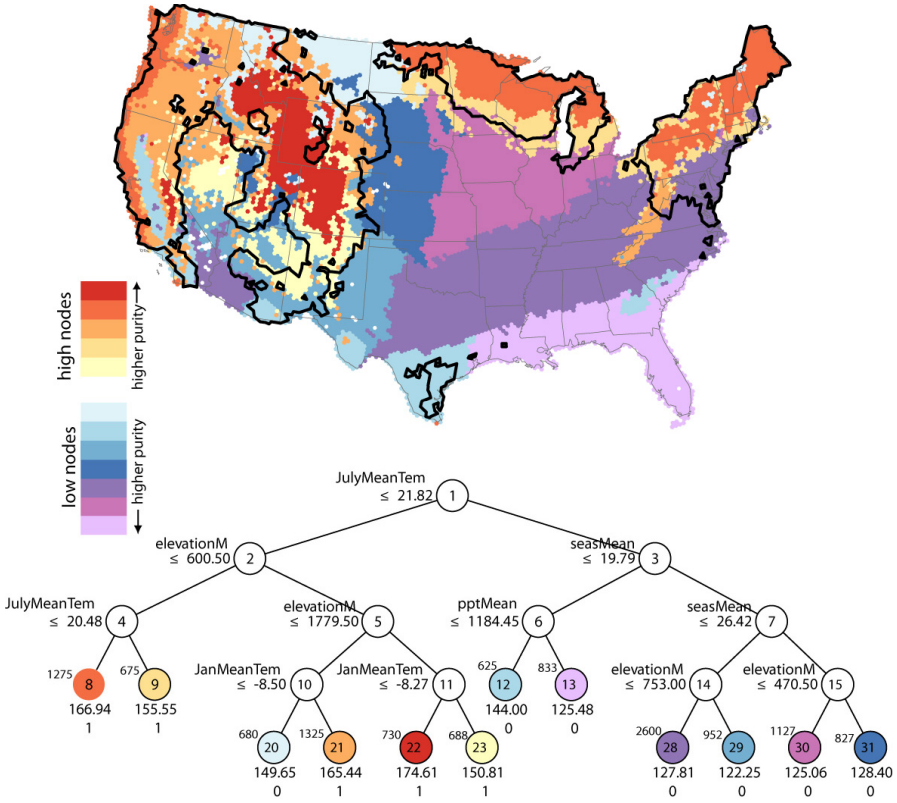

Figure 2. Map of richness of bird species environmental regimes calculated using regression tree. Tree nodes are shown as circles with ID numbers within them. Quantities on the left site of terminal nodes give the number of hexes in the node, quantities immediately below terminal nodes give an average value of *R* in the node, and the 0/1 labels indicate whether node represents *HR* regime or *LR* regime.

Results of classification tree model (*CTM*) are shown on Fig.3. The overall accuracy of the *CTM* is ~85%. The major split of dataset is on the value of January mean temperature, but *HR* and *LR* nodes are split between the two main forks of the tree. The *HR* node #9 accounts for majority of *HR* hexes.

4. Conclusions

The two models represent different means of decision tree learning and yield seemingly different partitionings of the dataset. From a prediction point of view they are equally useful although the *CTM* has a small edge in accuracy. From a point of view of discovering environmental regimes of biodiversity, each model provides what, at first glance, appears to be a different partitioning of the environmental data. However, closer examination reveals some similarities in spatial extent between a number of nodes in the two partitions. For example, spatial footprint of node #28 in the *CTM* resembles the footprint of node #12 in the *RTM*. Other examples include: *CTM* node #17 and *RTM* node #20, eastern portion of *CTM* node #9 and *RTM* node #8. These correspondences exist because a tree node is described in terms of a series of consecutive predicates, but a similar partition can be feasibly described by a different series of predicates if the predictor variables involved in the predicates are correlated.

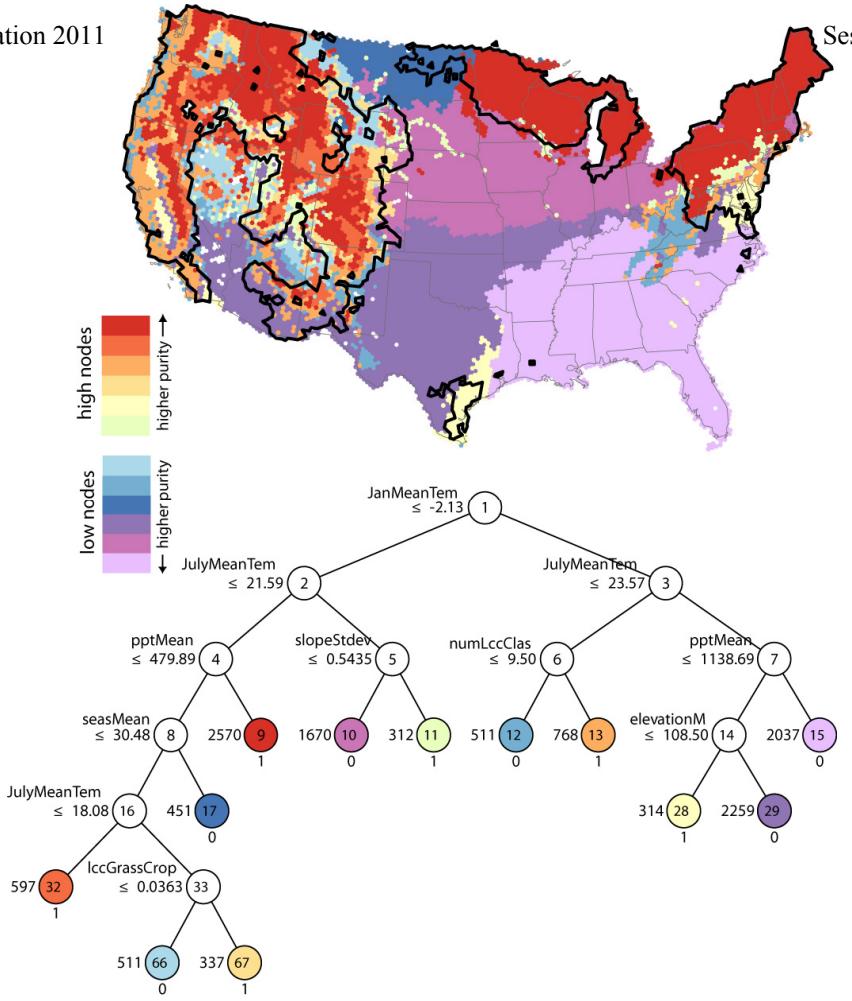

Figure 3. Map of richness of bird species environmental regimes calculated using classification tree. See also caption to Fig. 2.

Analysis of the two models reveals existence of four regimes of high richness of bird species that transcend specificity of the models. They are: (1) Southern regime (*RTM* node #12 and *CTM* node #28), (2) Northern regime (*RTM* node #9 plus portions of node #8 and *RTM* portion of node #9), (3) Mountain regime (*RTM* nodes #22, #21 and *CTM* nodes #32, portion of #9), (4) Pacific Coast regime (*RTM* portion of node #8 and *CTM* node #13). Fig.4 shows spatial extents of these regimes and their characterization in terms of predictors shown as parallel coordinates-like graphs. These characterizations provide compact but comprehensive description of each regime. For example, the Southern regime is not only characterized by climatic variables, as indicated by predicates in both regression and classification trees, but also by presence (predictor 17) and absence (predictors 18 and 19) of specific land cover classes.

Decision tree-based methodology, as presented here, can be applied to a broad range of non-stationary spatial problems where there is a need to identify different regimes of dependence between predictors and response.

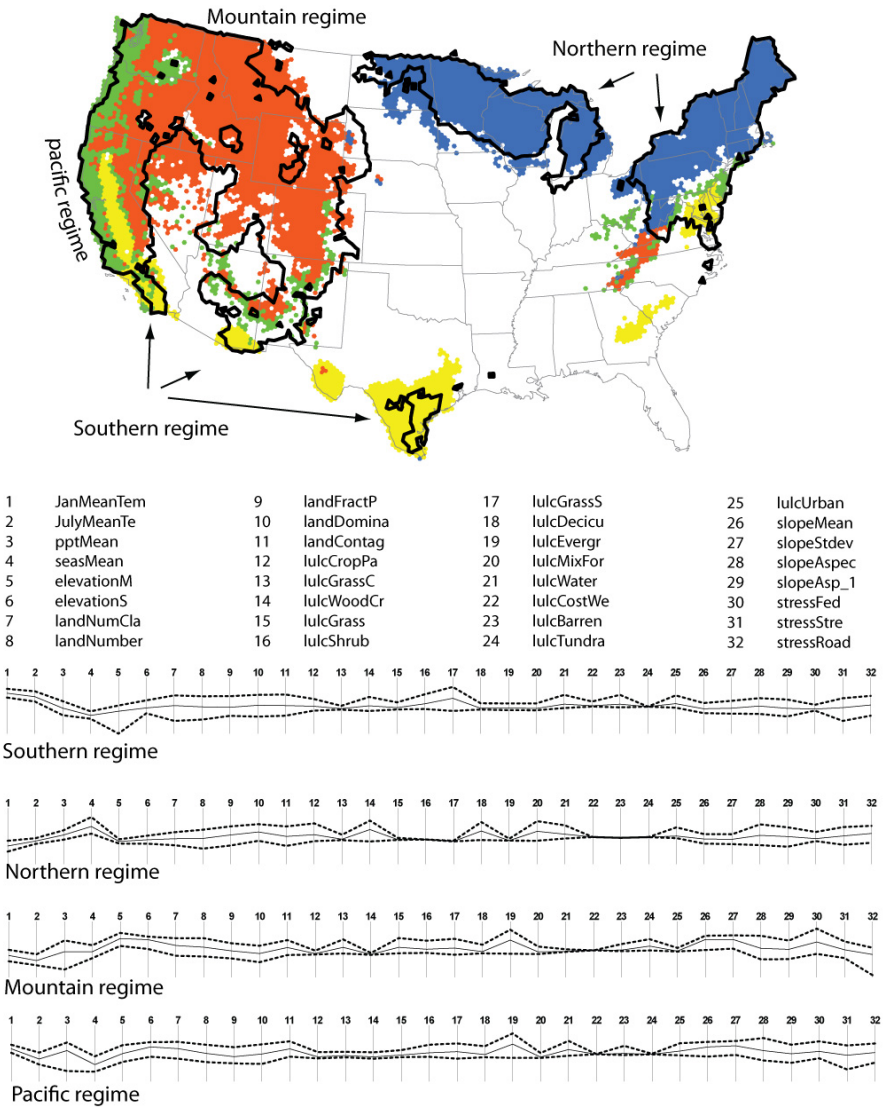

Figure 4. (Top) Map of four regimes of high diversity of bird species in the United States. (Bottom) Characterization of the four regimes in terms of all predictors; solid line – mean values, dashed lines – mean values ±standard variation.

7. References

Loh, W.-Y., 2008, Regression by parts: Fitting visually interpretable models with GUIDE. In: *Handbook of Computational Statistics, vol. III*, Springer-Verlag, 447-469.

Sauer, JR., Pendleton GW, and Orsillo, S, 1995. Mapping of bird distributions from point count surveys. In: Ralph CJ, Sauer JR, and Droege S (eds) *Monitoring Bird Populations by Point Counts, USDA Forest Service, Pacific Southwest Research Station,* General Technical Report PSW-GTR-149. 151-160.

White D, Preston B, Freemark K, and Kiester A, 1999, A hierarchical framework for conserving biodiversity. In: Klopatek J and Gardner R (eds), *Landscape ecological analysis: issues and applications*, New York: Springer-Verlag, 127-153.

White D and Sifneos J.C., 2002, Regression tree cartography. *J. Computational and Graphical Statistics*, 11(3):600-614.