

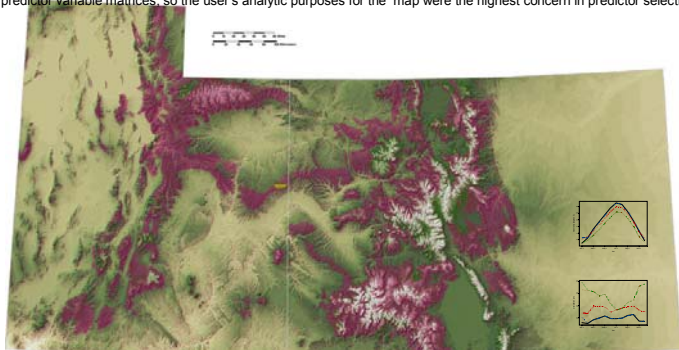
Quaking aspen biogeography: issues of fuzzy vs. probabilistic landcover mapping and multispectral vs. environmental predictors.

Joe Sexton, M.S. student, College of Natural Resources, Utah State University

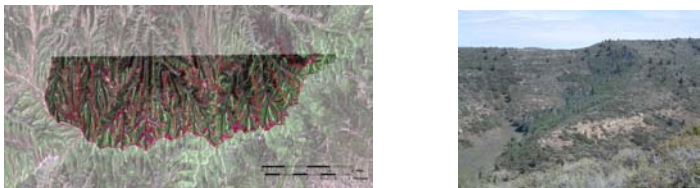
ABSTRACT

A currently popular tactic of predictive landcover mapping based on fuzzy set theory is to train a probabilistic statistical model on a matrix of one categorical landcover response variable and some combination of environmental and remotely-sensed predictor variables to output continuous predictions of subpixel composition. One implicit and rarely validated assumption in this procedure is the equality of predicted probabilities and sub-pixel coverage rates (interpreted as fuzzy-set class memberships) of the predicted classes. This assumption is uniquely true for each instance and few heuristics have been developed to predict its validity. Also, an implicit and often neglected assumption in the map's subsequent use as a predictor is that those analyses which make use of it are simpler than if they had merely used the map's own predictors. To confront these issues in mapping quaking aspen (*Populus tremuloides*, Michx.) for an analysis of its local decline, I explored the validity of the membership-probability assumption for a fuzzy, two-class raster grid of aspen modeled on climatic, topographic, and multispectral variables using logistic regression in a fuzzy classification framework.

Field-mapped aspen stand polygons of a 10,450.44-ha landscape in the Book Cliffs of eastern Utah, USA were rasterized to a resolution of 10m. This grid was then aggregated to 30m resolution, assigning each coarse (30m) grid cell with the proportion of its component 10m cells coded with aspen presence. Environmental, multispectral, and composite datasets were used to predict probabilities of 100% aspen cover from an endmember (i.e., consisting only of "pure" pixels) sample of aspen presence/absence using logistic regression. For both the training dataset and a separate but sympatric test dataset, correlation coefficients between the predicted probabilities and the observed percent cover of aspen showed that only modest agreement exists between probability of 100% aspen cover and fuzzy class membership. No large differences in predictive accuracy or probability-membership equality were found between environmental, multispectral, and composite predictor variable matrices, so the user's analytic purposes for the map were the highest concern in predictor selection.



The predicted geographic range of quaking aspen habitat in Utah and Colorado. The study area, located on the narrow east-west band of predicted habitat, is highlighted in yellow near the center of the image. Quaking aspen is the most widely-distributed tree species in North America. However, recent studies have concluded that Utah has lost 51% and Colorado has lost 49% of aspen cover since European settlement due to the interaction of its life history strategy with climate change, fire suppression, and overgrazing by wild and domestic ungulates. Economic and social values lost to this type-conversion have prompted managers to seek spatially-explicit analyses of its decline. Image created by Rob Johnson and Joe Sexton, USU/BLM LEMA Center, by superimposing rasterized NRCS STATSGO Plants Database polygons on a hillshaded and colorized 90m DEM mosaic. Insets compare monthly temperature (above) and rainfall (below) patterns of the study area and surrounding ecoregions (PRISM). Ecoregions in which aspen habitat is present have lower temperatures and higher rainfall in the growing season than surrounding areas, supporting a coarse-scale climatic limit to aspen's biogeographic range.



Study area and field survey of aspen stands. The steep-sided, parallel canyons and spatially discrete aspen stands of the Book Cliffs made it possible to conduct a full-coverage ground survey of aspen from walked transects. The mapped binary polygons were rasterized to a resolution of 10m and aggregated to 30m, recording in each 30m cell the percent of included 10m cells that were coded with aspen presence. The left image was created by Rob Johnson and Joe Sexton, USU LEMA Center, by superimposing transect lines (red) and rasterized aspen polygons (magenta) on a summer LANDSAT ETM natural-color image that was then differentially shaded to highlight the study area boundary. The photo on the right was taken by Joe Sexton from a survey transect. The distinct aspen stand has a general N-NE aspect and is being invaded by Douglas-fir (*Pseudotsuga menziesii*). Other spatially dominant woody vegetation in the area includes sagebrush (*Artemisia* spp.) and mixtures of Gambell's oak (*Quercus gambellii*), mountain-mahogany (*Cercocarpus* spp.), and serviceberry (*Amelanchier* spp.).

Environmental

Modeled monthly sums of potential evapotranspiration and direct shortwave radiation.

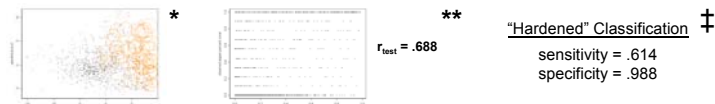


Called "ancillary data" by remote sensors, ecologically relevant data are used to produce **habitat prediction** maps. Environmental variables are assumed to be **necessary** cause for observed landcover, but predictions can only be interpreted as **potential** landcover maps due to the practical impossibility of capturing **sufficient** cause in the dataset. Further, **landcover** predictions may be extrapolated from **habitat** prediction maps only with additional information on the state of the system—a role to which remotely sensed data is currently applied. Because ecological relationships are most directly interpreted through the model (coefficients and p-values in the case of logistic regression), choice of model is limited by interpretability.

Multispectral

Scene: Landsat ETM+ scene 36/33

Bands: spring (April 2000): TM(1:6), TC(1:3)
summer (June 2000): TM(1:6), TC(1:3), NDVI
fall (November 1999): TM(1:6), TC(1:3), NDVI



Along with other remotely sensed data, multispectral data can be used directly to produce **landcover prediction** maps. In contrast to the predictor-response causal relationship between environmental and landcover data, the observed landcover is assumed to be **necessary** cause for the values of the predictor variables. The model itself is not directly ecologically interpretable, but *in cases where the spatial ecological relationships are of special concern*, these relationships may be interpreted between the predicted map and other spatial datasets (with, e.g., GIS overlay analysis, parametric and nonparametric statistics, simulation, etc.). Where nonspatial ecological relationships are to be interpreted, the original ground-truth or otherwise measured (i.e., not predicted) landcover data are preferable for their lack of uncertainty.

Composite

environmental + multispectral datasets



Combined datasets of environmental and remotely sensed variables are often used to produce landcover prediction maps with higher accuracy than that provided by environmental or remotely sensed data alone. However, despite improved potential accuracy, robust ecological interpretation is hampered by confounding between environmental and remotely sensed variables. Also, ecological inferences made from the interpreted map may be circular, and it is recommended that predictor variables be chosen not only for their prediction value, but also for their **independence from relationships to be inferred from the predicted map**.

- * Scatterplots of 1000 randomly sampled 30m pixels on 1st and 2nd principal components of (from top to bottom) environmental, multispectral, and composite datasets. Black points show aspen absence, and colored circles of increasing diameter show increasing aspen percent cover at 30m resolution (aggregated from 10m resolution pixels).
- ** Assessment of models on separate, sympatric test datasets: probability-membership correlation. Scatterplots of 1000 randomly sampled pixels of (from top to bottom) environmental, multispectral, and composite datasets, showing the correlation between a cell's degree of membership in class "aspen" and its predicted probability of belonging completely to class "aspen" (to a degree of 1.0). All logistic regressions were optimized with backward stepwise variable selection. The banding is an artifact of the small difference between cell sizes in the aggregation step (9 cells at 10m resolution aggregated to 1 cell at 30m resolution).
- ‡ Assessment of models on separate, sympatric test datasets: prediction accuracy for discretized logistic regression. Predictions were "hardened" at a cutoff of p(ASPEN) = 0.5. Sensitivity is the proportion of correctly classified aspen presences, and specificity is the proportion of correctly classified aspen absences. Prior probability of aspen in the training dataset was .0455. The apparent superiority of the multispectral classification is probably an artifact of stepwise variable selection.

QUESTIONS (your input is appreciated)

•Could accuracy be improved by creating a hierarchical model of geographically smoothed/resampled variables?



•Would a more flexible classification (e.g., QDA, CART) or regression algorithm (e.g., GAM, MARS) provide better accuracy and/or probability-membership equality? (Incorporation of fuzzy set-theoretic axioms may be helpful for generalizing regression to a k-dimensional response variable/vector, *but how?*)

•How does one sample "independent" observations of a spatially and temporally autocorrelated dataset? What about one with a binary response variable?



The agreement between the original "ground truth" map and the predicted map of aspen using the composite dataset. Green polygons are the ground-truth aspen stands, and increasing red saturation corresponds to increasing probability of binary aspen presence. The zoomed inset corresponds roughly to the area inside the black box on the full coverage grid.