Using ArcView in Conjunction with Statistical Software for Predictive Modeling

Chris Garrard

Predictive spatial modeling in ecology is becoming more common, but unfortunately, not all ecologists possess the skills necessary to successfully combine statistical models and geographical information systems. In response to this problem, I wrote an ArcView extension which provides an interface between ArcView and the SAS and S-PLUS statistical software packages. The extension provides a graphical user interface which helps users create data sets with which to build models, creates script files containing the necessary code for the statistical software, and imports and maps the model results in ArcView. Logistic regression and classification and regression tree models are supported.

Introduction

As geographical information systems (GIS) become more accessible and GIS data become more plentiful, statistical models within a spatial context are showing up more often in the ecological and conservation literature. For example, logistic regression models have been used with GIS to create habitat models for the fisher (Carroll et al. 1999) and the Eurasian lynx (Schadt et al. 2002). Classification and regression tree (CART) models have been used with GIS to produce habitat suitability models for smallmouth bass (Rejwan et al. 1999) and desert tortoises (Andersen et al. 2000).

Mapping the predictions of statistical models over a spatial area is not always straightforward, however, especially if the modeler has minimal knowledge of GIS. The process requires at least a basic understanding of projections, resolution, and data formats. It also requires the user to be able to transfer data between the GIS and statistical software, and then implement the model results in the GIS. These tasks can be very daunting for the average user. In addition, some types of models, such as CART, are time consuming to implement even for the experienced user. As predictive spatial modeling in ecology becomes more common, the need arises for more tools to help with this modeling process.

StatMod: A Tool for Statistical Modeling with ArcView GIS

To simplify spatial modeling tasks for ecologists, I developed an ArcView GIS® extension, called StatMod, which interfaces ArcView with SAS/STAT® and S-PLUS® statistical software. The extension provides a graphical user interface (GUI) to help the user perform logistic regression modeling using SAS and CART modeling using S-PLUS. StatMod helps the user sample ArcView themes, set up analysis options, run the statistical analysis, and read the resulting model back into ArcView and display the results as a map. StatMod also contains routines to help the user convert and resample data, select random points, and perform Kappa analyses. A list of functions available in StatMod is shown in Figure 1.

In order to use StatMod, the ArcView and the appropriate statistical package must be available. StatMod does not include any of these software programs. This extension is freely available from www.gis.usu.edu/~chrisg/avext/.

StatMod Menu Option	Function
Logistic Regression (SAS)...	Parameterize a logistic regression model
Tree Models (S-PLUS)...	Parameterize a CART model
Import Model...	Import StatMod output from SAS or S-PLUS
Specify Model...	Import a model not created with StatMod
Edit Tree...	Manually edit the rules for a tree
Create Rules...	Create an easy-to-read listing of rules included in a tree
Sample...	Sample themes at locations contained in a point theme
Zone Sample...	Create summary statistics for grid cells contained within the boundaries of polygons
Random Sample...	Create a random sample of points
Compute Kappa...	Estimate Cohen's Kappa and create an error matrix for a model
Intersect Themes...	Intersect polygon themes
Convert to Grid...	Convert a polygon theme to a grid theme
Resample...	Resample a grid theme to a new cell size
Combine Grids...	Merge adjacent grid themes
Theme Information	Display model information for a theme created with StatMod
Properties...	Set StatMod properties
Help	Show StatMod help
About StatMod...	Show StatMod copyright statement

Figure 1. StatMod menu options and functions.

Logistic Regression

Logistic regression models attempt to predict the conditional probability of an event occurring, given the values of independent variables (Hosmer and Lemeshow 1989). For example, if logistic regression analysis were used to model presence of bobcats, then the model output would be probabilities of bobcats occurring under different conditions. StatMod incorporates many of the options available in SAS for logistic regression. Most of the SAS options that are not available in the StatMod GUI can still be used, but with a little more work on the user's part. The main logistic regression dialog is shown in Figure 2, and the options available are listed below.

General (Figure 2)

- Both SAS single-trial and event/trial syntaxes are supported.
- Dependent variables may come from any string or numeric field in an attribute table for a point theme.
- Independent variables may come from any string or numeric field in an attribute table for a polygon or grid theme. They may also come from floating point grids.
- Independent variables can be designated as class (factor) variables.
- Two-way interactions are supported, although higher order interactions are not.
- A table containing the data used to build the model can be added to the current project.

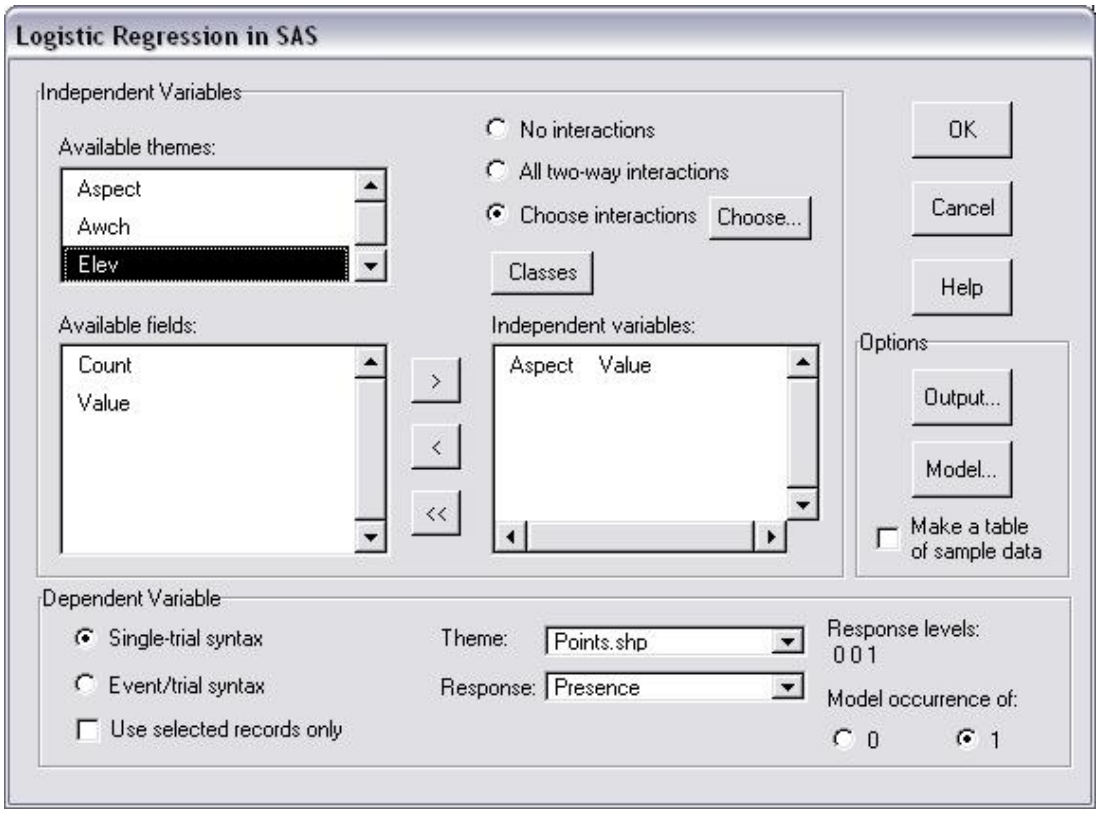


Figure 2. Main logistic regression dialog.

Printed Output

- Simple descriptive statistics for each independent variable in the model.
- Detailed results from each iterative step in the model fitting process.
- R^2 value for the fitted model.

- Correlation and covariance matrices of parameter estimates.
- Confidence intervals for parameters and odds ratios.
- Hosmer and Lemeshow goodness-of-fit test for binary response models.
- Pregibon's diagnostic measures for identifying influential observations.

Model Options

- Choice of backward, forward, stepwise, or no variable selection method.
- Can set minimum or maximum number of variables included in the final model.
- Can set the maximum number of times a variable can be added or removed from the model during the fitting process.
- Can set significance levels for adding and removing variables during the fitting process.
- Can set maximum number if iterations of the fitting process.
- Model intercept is optional.
- Certain variable can be forced into the model.

Classification and Regression Trees

Classification and regression tree (CART) models are nonparametric, can use several types of independent variables, and are not significantly affected by outliers (Breiman et al. 1984, Verbyla 1987). Classification trees predict discrete classifications while regression trees provide estimates of the dependent variable. In both cases, the structure of the model is that of a binary tree; the terminal nodes constitute the predictions. StatMod contains some algorithms for determining the size of the fitted tree, or the user can fit the tree interactively in S-PLUS. The main CART dialog is shown in Figure 3, and the options available are listed below.

General (Figure 3)

- Point Sampling
  - Dependent variables may come from any string or numeric field in an attribute table for a point theme.
  - Independent variables may come from any string or numeric field in an attribute table for a polygon or grid theme. They may also come from floating point grids.
- Zone Sampling
  - Dependent variables may come from any string or numeric field in an attribute table for a polygon theme.
  - Independent variables may come from grid. The value of the variable is a summary statistic of grid values that fall inside the polygon. Available statistics include minimum, maximum, majority, minority, variety, mean, median, standard deviation, sum, and range.

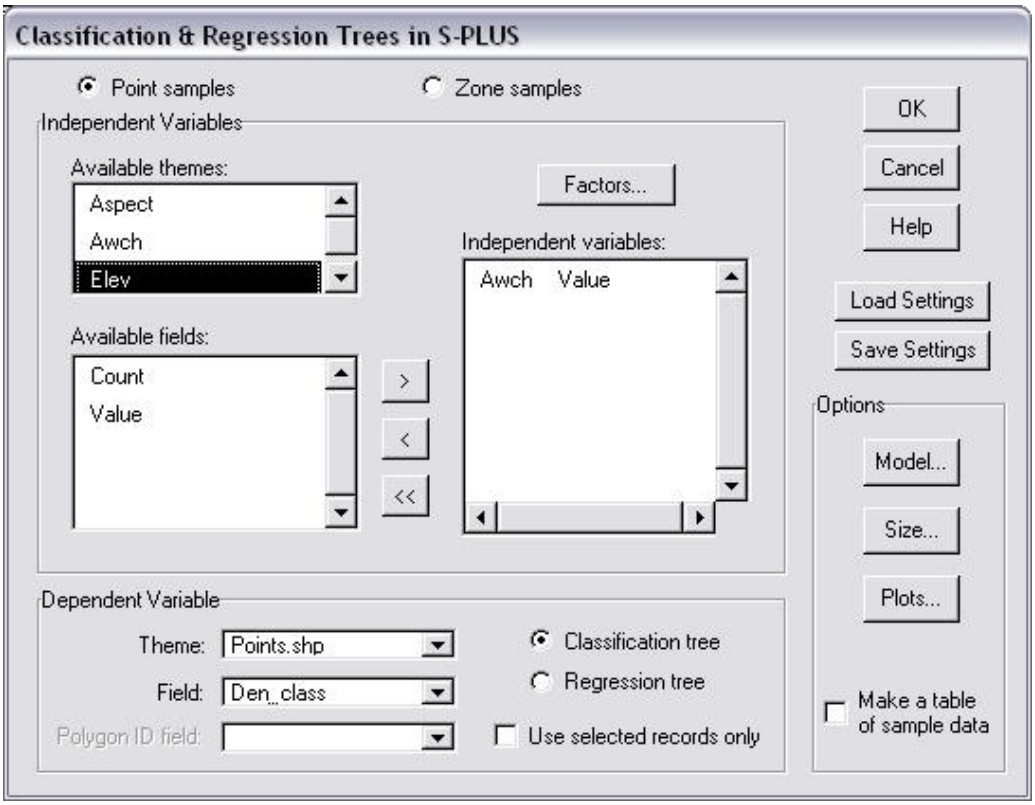


Figure 3. Main classification and regression tree dialog.

- Independent variables can be designated as factor variables.
- A table containing the data used to build the model can be added to the current project.
- Model settings may be saved to a configuration file and loaded again at a later date.

Model Options

- If not fitting a full tree or using S-PLUS default parameters, the following parameters may be set:
  - The number of observations that a node must contain before it is eligible for splitting.
  - The minimum number of observations allowed in a node.
  - The minimum allowed deviance for a node.

- Records with missing values for some of the variables can either be excluded from the analysis or, if the missing values are for factor variables, a new factor level for missing data can be created.

Tree Size Options (Figure 4)

- The same data that was used to grow the over-fitted tree may be used to evaluate subtrees, or a new data set may be designated.
- The tree may be pruned using deviance or misclassification rate to evaluate the subtrees, or the tree my be shrunken.
- The following methods for determining tree size may be used (or the tree may be pruned interactively in S-PLUS):
  - One standard error rule (with or without cross-validation).
  - Akaike's Information Criterion (classification trees only).
  - Mallow's C_p (regression trees only).
  - A specific tree size may be returned.
  - A pruned tree with a specific cost-complexity parameter may be returned.
  - A shrunken tree with a specific shrinkage parameter may be returned.

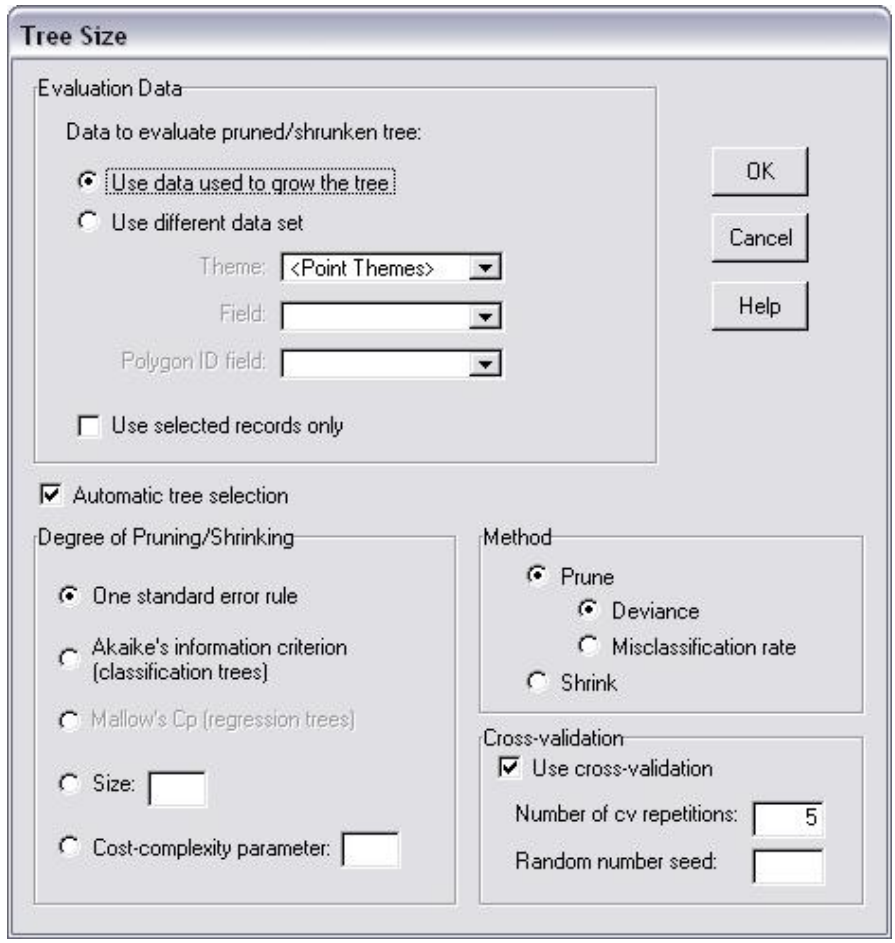


Figure 4. Tree Size options dialog.

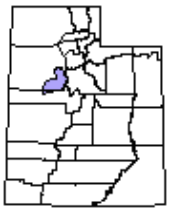
Plot Options

- Plots may be shown in S-PLUS or saved to a PDF file.
- Graphical plot of the fitted tree. Branch length may be uniform or based on deviance.
- Plot of deviance vs. tree size.
- Normal probability plots of original and fitted trees (regression trees only).
- Plots of residuals vs. predictions for original and fitted trees (regression trees only).

Kappa Analysis

Error matrices and Cohen's kappa (κ) are commonly used for accuracy assessment. Possible applications include accuracy assessment for models that predict discrete classifications or classifying imagery. Kappa can be used as a measure of agreement between model predictions and reality (Congalton 1991) or to determine if the values contained in an error matrix represent a result significantly better than random (Jensen 1996).

The Kappa Analysis tool in StatMod creates error matrices and computes κ. In addition, the standard error around κ and the associated *z*-score are also computed. The *z*-score can be used in conjunction with a table of critical values for the standard normal distribution to determine statistical significance. Kappa is considered to be a reliable measure of accuracy if the *z*-score is significant (Fleiss 1973).



Mapping Sage Grouse Habitat

In an effort to delineate sage grouse habitat, the RS/GIS Lab at Utah State University mapped landcover in four Utah study areas (RS/GIS Laboratories 2003). The result of one of these study sites, Rush Valley, will be discussed briefly here.

Data were collected at a total of 208 sites (about 4,000 acres) over a two month period. Of these, 168 were used to build a land cover model using classification trees. The remaining 40 sites were used for accuracy assessment.

A total of 13 independent variables were input as independent variables in the model. These consisted of the three tassel-cap bands (brightness, wetness, and greenness) for three different Landsat ETM images (spring, summer, and fall dates), elevation, slope, aspect, and landform. These were sampled using the Zone feature of StatMod and a tree was generated. The mapped predictions are shown in Figure 5.

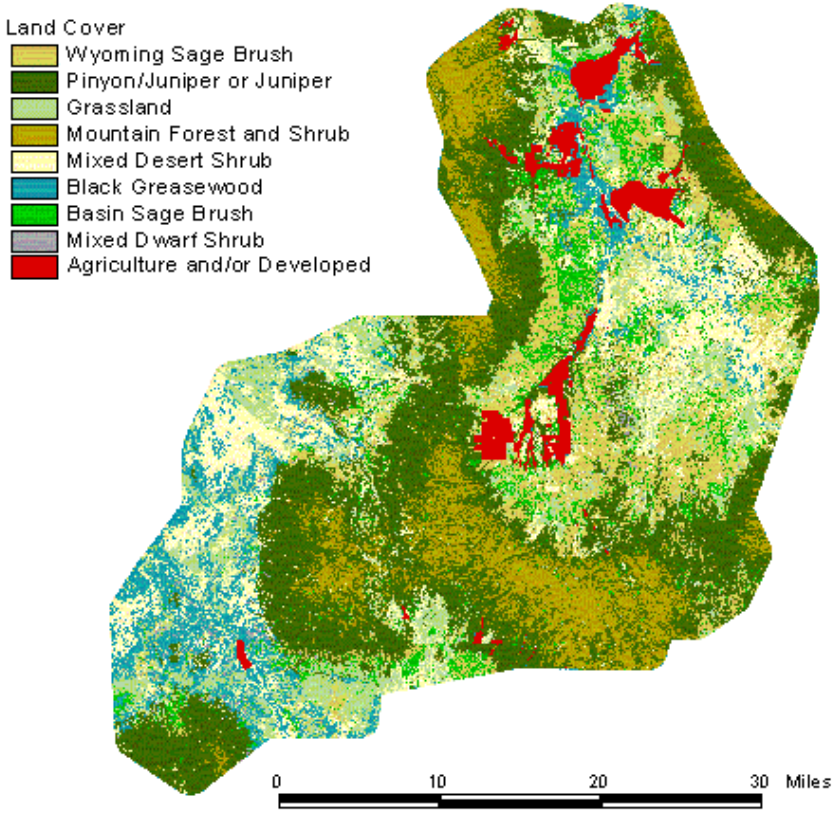


Figure 5. Land cover predictions for Rush Valley, UT.

Overall accuracy and κ were computed for this predictive model using the 40 sites that were held out for accuracy assessment. The overall accuracy was 0.70 and κ was 0.63.

A regression tree analysis was then done in an attempt to model sagebrush canopy cover. All non-sagebrush cover types were masked out of the analysis, so that the classifier was only attempting to model canopy cover where sagebrush was dominant. The results of this exercise are not as reliable as those from the first model, but they are potentially useful. A detailed view of the results for a small portion of the study site is shown in Figure 6.

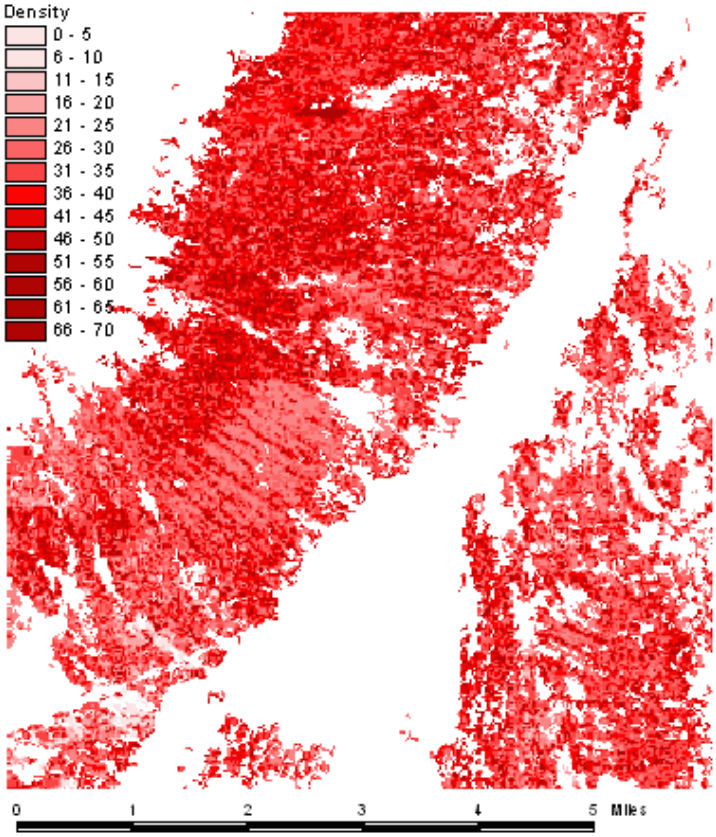


Figure 6. Sagebrush canopy cover predictions for Rush Valley, UT.

Mapping Vegetation in Grazing Allotments

A rancher in Squaw Valley, NV, wanted more information to base his grazing strategy on. Members of Bear River Geospatial used StatMod to help provide this information.

First, they put together some data layers for the grazing allotments. These included elevation, slope, aspect, landform, multiple dates of Landsat imagery, and indices derived from the satellite images. They then loaded these data sets onto a laptop that also had ArcView, StatMod, and S-PLUS, and headed out into the field.

Days were spent collecting data, and these data were used to build a classification tree model in the evenings. From their knowledge of the area, the researchers could evaluate the models and determine what vegetation classes were not being predicted accurately. This helped them focus their data collection for the following day. For example, the first models did not delineate the wet meadows within the allotment, so they collected data at more wet meadow sites so that this cover type was predicted by subsequent models. The final model (Figure 7) had an overall accuracy of approximately 80%.

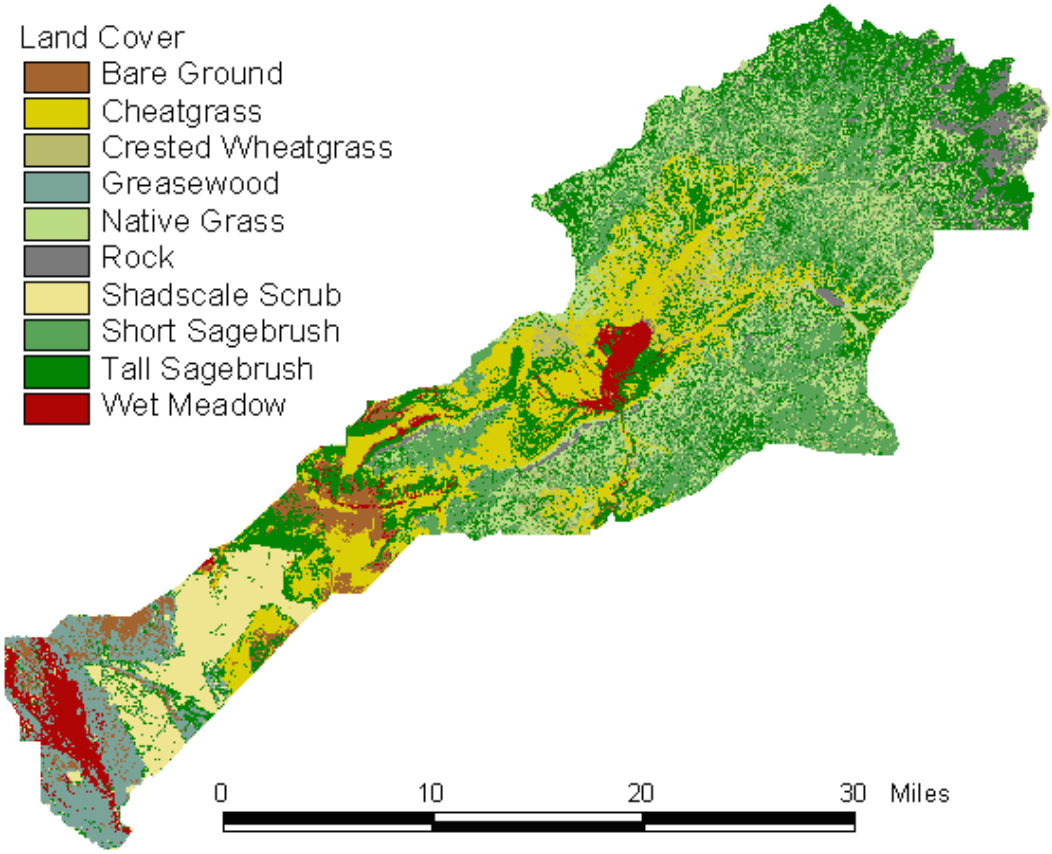


Figure 7. Predicted vegetation classifications in Squaw Valley, NV.

A graphical representation of the tree used to produce Figure 7 is shown in Figure 8. Only the first split condition is shown, but the figure still shows the size and complexity of the model.

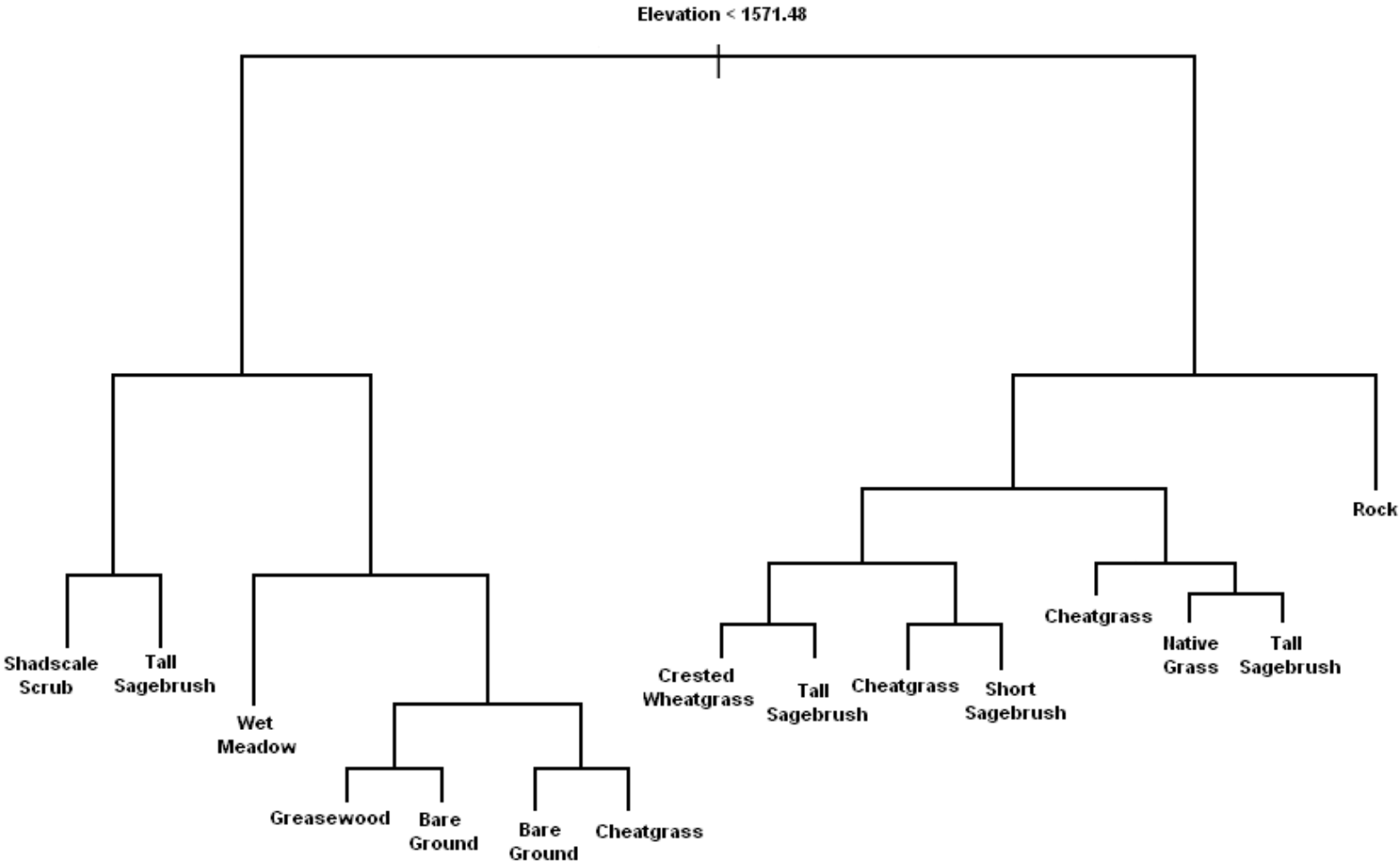


Figure 8. Classification tree model corresponding to the mapped predictions shown in Figure 7.

Acknowledgments

Development of StatMod was originally part of my thesis project in the Department of Biology at Utah State University. Dr. Jim Haefner advised this first (and major) section of the development. Later work was done under the employment of the RS/GIS Laboratories at Utah State University. John Lowry, Eric Sant, and Lisa Langs provided ideas and requests for features during this phase. Thanks to John Lowry, Eric Sant, and Chris McGinty for examples of their models.

References

Andersen, M. C., J. M. Watts, J. E. Freilich, S. R. Yool, G. I. Wakefield, J. F. McCauley, and P. B. Fahnestock. 2000. Regression-tree modeling of desert tortoise habitat in the central Mojave desert. *Ecological Applications* 10: 890-900.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman & Hall, New York, New York, U.S.A.

Carroll, C., W. J. Zielinski, and R. F. Noss. 1999. Using presence-absence data to build and test spatial habitat models for the fisher in the Klamath region, U.S.A. *Conservation Biology* 13: 1344-1359.

Congalton, R. G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37:35-46.

Fleiss, J. L. 1973. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, Inc., New York, New York, USA.

Hosmer, D. W., and S. Lemeshow. 1989. *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, New York, USA.

Jensen, J. R. 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective* (Second edition). Prentice Hall, Inc., Upper Saddle River, New Jersey, USA.

Rejwan, C., N. C. Collins, L. J. Brunner, B. J. Shuter, and M. S. Ridgway. 1999. Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* 80: 341-348.

RS/GIS Laboratories, Utah State University. 2003. *Sagebrush Ecosystem Mapping using Landsat ETM, Final Report*. Unpublished manuscript.

Schadt, S. E. Revilla, T. Wiegand, F. Knauer, P. Kaczensky, U. Breitenmoser, L. Bufka, J. •ervený, P. Koubek, T. Huber, C. Staniša, and L. Trepl. 2002. Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology* 39: 189-203.

Verbyla, D. L. 1987. Classification trees: a new discrimination tool. *Canadian Journal of Forest Research* 17: 1150-1152.

Author Information

Chris Garrard, Research Associate
RS/GIS Laboratories, Utah State University
5275 Old Main Hill
Logan, UT 84322-5275
435-797-2602
chrisg@gis.usu.edu