TECHNIQUE FOR RANKING POTENTIAL PREDICTOR LAYERS FOR USE IN REMOTE SENSING ANALYSIS

Andrew Lister, Mike Hoppus, and Rachel Riemann

ABSTRACT. Spatial modeling using GIS-based predictor layers often requires that extraneous predictors be culled before conducting analysis. In some cases, using extraneous predictor layers might improve model accuracy but at the expense of increasing complexity and interpretability. In other cases, using extraneous layers can dilute the relationship between predictors and target variables that the modeling technique seeks to exploit. The current study seeks an automated method whereby a ranking of potential predictor data can be obtained so that the researcher can quantitatively justify including or excluding certain predictors. In our example, we seek to determine the relative strength of the relationship between various Landsat ETM satellite spectral layers and total basal area on 380 study plots established by the USDA Forest Service's Forest Inventory and Analysis unit.

KEYWORDS. Forest inventory, Landsat, multiple regression

Introduction

The USDA Forest Service's Forest Inventory and Analysis unit (FIA) collects information from ground plots, summarizes this information, and produces statistical and analytical reports that document the quantity, distribution, composition and health of the nation's forests. In order to accomplish these tasks, FIA uses spatial modeling methods. Many of these spatial modeling techniques require that both the target variables (e.g., basal area) and the predictor dataset (e.g., satellite images, GIS data) be prepared before analysis by culling extraneous data and eliminating outliers.

Several steps have been used to eliminate extraneous predictor data and reduce collinearity among the predictors. For example, summary statistical analyses and scatterplot analysis have been used to identify high correlations among predictor data and weak correlations between predictor and target data (Lister *et al. In Press*). This process is cumbersome when there are numerous predictor layers, and is difficult to perform objectively. Also, simple bivariate scatterplots or correlation coefficients do not reveal information about multidimensional relationships; certain relationships in the data might be mediated by the presence of other data in the model.

Principal components analysis has been used to reduce data redundancy by creating new variables from a linear combination of input layers (Lister *et al. In Press*; McGarigal *et al.* 2000). These new variables are orthogonal to each other and describe axes of maximum variation of the predictor dataset as a whole. A disadvantage of using this approach is that these new variables only incorporate information on the suite of predictor data layers but not information on the relationship between the predictors and the target data. Furthermore, it is difficult to attach biological meaning to the principal components.

Linear modeling approaches, for example canonical correlation analysis or multiple regression, incorporate information on the relationships between the predictor and target data. While subject to various limitations described elsewhere (e.g., Montgomery and Peck 1982), a regression-based approach is a convenient and easily understood method for identifying predictor variables that help describe variation in the target data. Traditionally, stepwise, forward and backward selection methods have been used to build these linear models. However, with such approaches, as a model is built, the inclusion of a certain variable might preclude the later inclusion of another predictor variable with a strong relationship with the target data. In some cases this is desirable, as including collinear predictors in a regression model not only affects the stability of the parameter estimates (Draper and Smith 1981), but needlessly adds to the complexity of the model and influences the interpretation of its standard error.

However, there are spatial modeling techniques for which highly correlated predictor variables might be beneficial. For example, in a regression tree context, two collinear variables might have different relationships within different regions of the target data's distribution, so it might be beneficial to make both variables accessible to the modeling procedure. For remote sensing, it is clearly advantageous to have a mature understanding of the multivariate interactions between predictor and target data. Selecting variables for inclusion in models based on traditional stepwise, forward, and backward selection methods severely limits one's understanding of these multivariate interactions because the typical result of these model building exercises is only one or a small number of models.

We recommend an approach that is an adaptation of the all-subsets method and is described by Shtatland *et al.* (2001). Every combination of predictor data layers is assessed in a model, and for each model, fit statistics and parameter estimates are stored as rows in a database. This database can be summarized to identify variables that appear most frequently in models deemed "good" with respect to a particular fit criterion. We present an example of the method using Landsat imagery and FIA plot data, and discuss the benefits and limitations of this approach.

Methods

Data were collected using methods outlined in USDA Forest Service (2000)[1]. The total basal area of all species of trees and saplings was computed on all 380 FIA plots that were 100-percent forested (Figure 1). The pixel values for three seasons of six band Landsat ETM+ imagery (spring, leaf on (l-on) and leaf off (l-off) bands 1-6) were obtained from the location beneath the center subplot of the FIA plots and assembled in a database that, along with basal-area values for the plots (the target variable), made up the modeling data set. Both the plot data and the Landsat images were acquired between 1999 and 2001. All variables' distributions were assessed visually for normality and transformations were applied as appropriate.

[1] USDA Forest Service. 2000. Forest inventory and analysis national core field guide, volume 1: Field data collection procedures for phase 2 plots, version 1.4. USDA Forest Service, Internal report. On file at USDA Forest Service, Washington Office, Forest Inventory and Analysis, Washington, D.C.

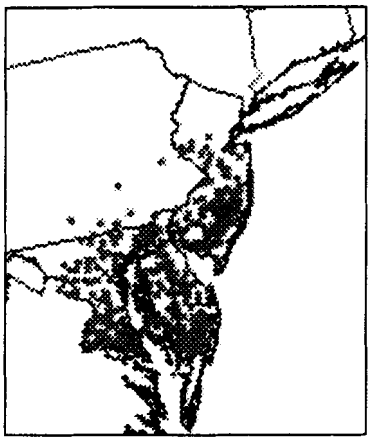

Figure 1. The 380 FIA study plots in New Jersey, Delaware, Maryland, and Pennsylvania.

Linear models of every combination of the 18 predictor layers (2^{18} or 262,144 combinations) were produced using ordinary least-squares regression (OLS) using SAS[2] statistical software (SAS Institute 1999). The parameter estimates and other information criteria (Akaike's Iinformation Criterion (AIC), R^2 adjusted for degrees of freedom of the model (R^2_{adj}), Mallow's Cp Index (Cp), Schwarz' Bayesian Information Criterion (SBC) and Root Mean Square Error (RMSE)) were computed using standard methods (SAS Institute 1999) and stored in a database in which each row represented information for one of the 262,144 models. This database was summarized both graphically and in tabular format to depict the relationships between number of variables in the model and the R^2_{adj}, as well as information on the frequency of occurrence of each of the 18 layers in models that were deemed to be relatively "good" ($R^2_{adj} > 0.3$) after the data summary was examined. We also determined the frequency of occurrence of each variable in models with 8 to 12 variables and with all parameter estimates that were significant at the 0.1 level. Forward (with all variables being forced into the model) and backward (with all but one variables being forced out) selection methods were then applied to the dataset for comparison purposes.

Results and Discussion

The frequency distributions of the R^2_{adj} values for each category of model complexity are depicted as a series of box and whisker plots in Figure 2. As more variables are included in the models, the mean R^2_{adj} follows an expected pattern of increasing and then leveling off with increasing model complexity. This occurs because certain predictor layers explain more of the variation in the target variable than others, and once that variation is explained by including those dominant predictors, smaller gains in explanatory power are gained with the addition of the less dominant predictors.

[2] The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the USDA or the Forest Service of any product or service to the exclusion of others that may be suitable.

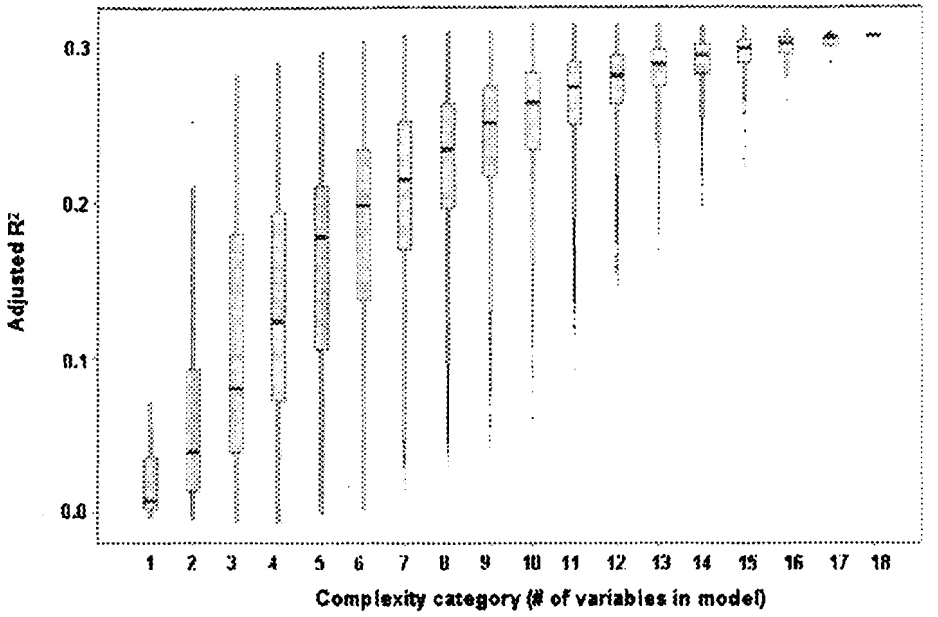

Figure 2. Relationship between complexity category and the distribution of R^2_{adj} values. The wide box represents the interquartile range (IQR) for each distribution. The length of the whiskers (thinner bars) is 1.5 times the IQR or the difference between the 25th (75th) percentile and the lowest (highest) value, whichever is smaller. Isolated dots represent values beyond the whiskers.

A useful feature of graphical depiction of model complexity vs. model fit is that one can quickly assess how different levels of complexity create a broad distribution of model-fit statistics. This suggests that a single model or a small number of models that we might have obtained resulting from any combination of frontward, backward or stepwise regression would have been inadequate if the goal was to understand the variability of model fit with respect to model complexity, and of model composition for a given level of goodness of fit.

Figure 3 shows how other information indices change along with increasing model complexity. On average, the other information indices (which were standardized for comparison) show the same pattern as R^2_{adj} -- beyond about eight variable models, the information criteria no longer show large changes, suggesting that the most efficient model contains fewer than eight variables. This information can help one construct an efficient linear model using these predictor data if this is the goal. Although Figure 2 shows the average information index of all models with a given number of variables included, the same model complexity vs. fit assessment can be made using stepwise procedures to help guide the choice of model.

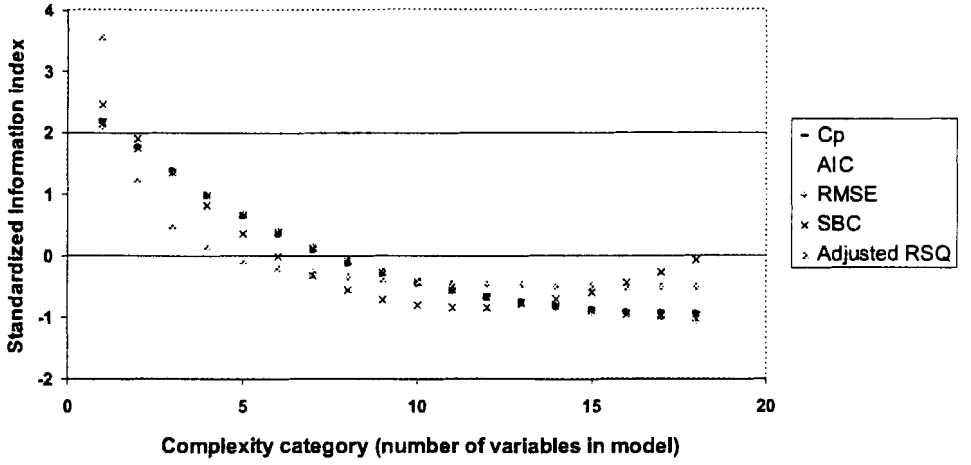

Figure 3. Relationship between model complexity and average values for several information criteria. Averages were computed across all possible combinations of models for each level of complexity.

Figure 4 depicts the frequency of inclusion of each of the 18 input variables in the "good" models. The relative heights of the different variables' bars generally remain constant across all levels of model complexity, indicating that the popular variables within the group of models with $R^2_{adj} > 0.3$ remain popular across different combinations of input variables -- their popularity is not an artifact of a specific configuration of input variables.

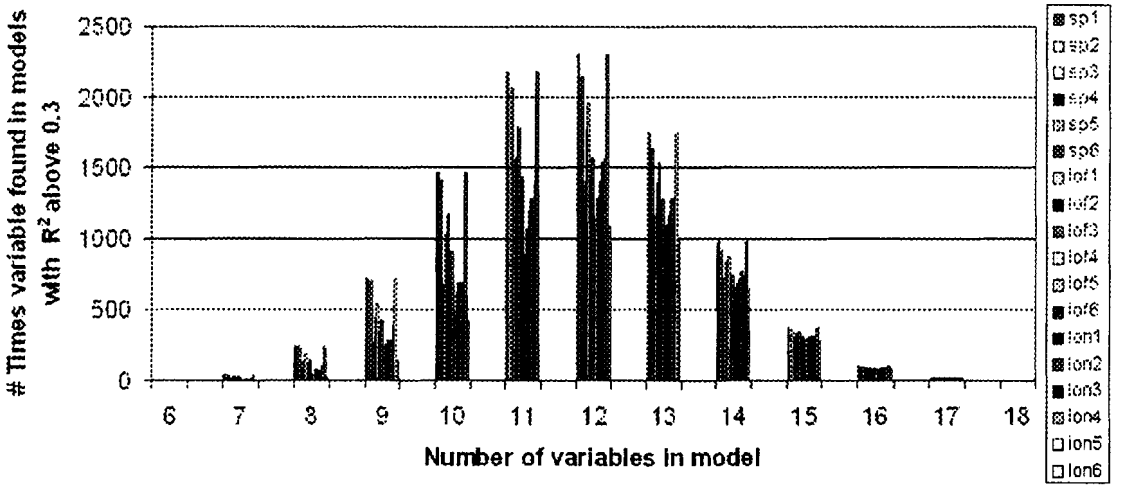

Figure 4. Frequency of inclusion of each of the input layers in the models with an $R^2_{adj} > 0.3$. The variable name convention is: sp (spring), lof (leaf off), lon (leaf on) – bands 1-6. Note that the relative heights of the bars generally remain the same across all categories of model size.

An assessment of Figure 2 led us to arbitrarily define a good model as one yielding an R^2_{adj} of 0.3 or greater. On the basis of our assessment of Figures 2 and 3, we chose to more closely

examine models with 8 to 12 variables. We call this the efficiency threshold, or the range of model complexity within which diminishing returns were obtained by adding extra variables. Figure 5 is an extract of the data from complexity categories 8 to 12, and is arranged to allow a quick visual assessment of the ranking of the popularity of variables included in the suite of good models and near the efficiency threshold. It shows all of the models (left axis), and only models for which all parameter estimates are significant at the 0.1 level (right axis). Table 1 shows the ranks of the variables from three variable ranking methods -- forward, backward, and all-subset (using only models with significant parameter estimates). The forward and backward methods ranked the variables in order of inclusion or duration of retention, respectively, in the model.

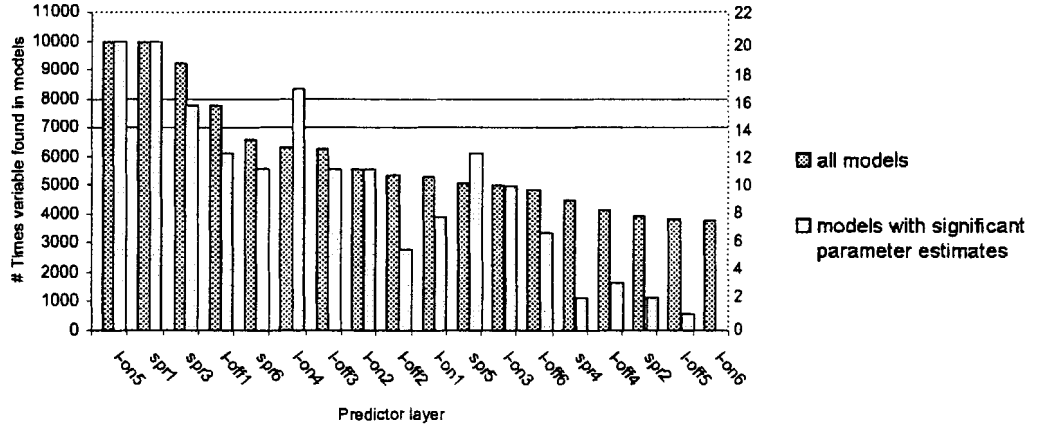

Figure 5. Frequency with which each input layer occurs in models with an $R^2_{adj} > 0.3$ and 8 to 12 variables. The variable name convention is: sp (spring), lof (leaf off), lon (leaf on) – bands 1-6. On the left axis, the scale represents the range of values for all models, and on the right side, the scale represents the range of values for models with all significant parameter estimates.

The methods yield similar results with a notable exception. Leaf-on band 6 (lon6) was ranked high in the forward selection process but low in the backward and all-subset methods. The forward selection process included l-on5 early in the process; since the Pearson's correlation between l-on5 and l-on6 is high ($r=0.81$, $p<.0001$), adding l-on6 after l-on5 explained only a small amount of extra variation in the dataset, so it appeared low in the ranking.

This situation illustrates the primary advantage of the all-subsets method -- one is not limited to a static ranking of variables with respect to their inclusion in a model via a stepwise selection procedure. Results from the latter procedure can be ambiguous and are highly influenced by the covariance structure of the predictor data layers (Draper and Smith 1981). Also, information on variables that nearly were included at each step generally is not available, nor is there a quantitative way to rank variables with respect to the frequency of their inclusion in a set of models with desirable characteristics.

Table 1. Ranking of the predictor variables in descending order of importance based on the all-subset (only models with significant parameter estimates), backward, and forward selection procedures.

All Subset	Backward	Forward
l-on5	l-on5	l-on6
spring1	spring1	spring1
l-on4	spring3	l-on5
spring3	l-off3	spring3
l-off1	l-off1	spring6
spring5	l-off2	l-off6
spring6	spring6	spring5
l-off3	spring5	l-on1
l-on2	l-on4	l-on2
l-on3	l-on3	l-on4
l-on1	l-on2	l-on3
l-off6	l-on1	l-off1
l-off2	l-off6	l-off2
l-off4	l-off4	l-off3
spring4	l-off5	l-off4
spring2	spring2	l-off5
l-off5	l-on6	spring2
l-on6	spring4	spring4

In our example, we considered a model as desirable if it had relatively high R^2_{adj} values. We also assessed the ranking of variables in this category and that were in models in which all of the parameter estimates were significant. We could have chosen other information criteria (such as those in Figure 3) to determine a good model, or we could have altered our procedure to define a good model as one that performs well in a cross validation. In either case, we were able to obtain a ranking of the variables, and also obtain frequency distributions of these "goodness criteria" for models with various levels of complexity. This additional information gave us a more mature understanding of the variability of some of the multivariate interactions that occur within the dataset, and might allow us to make decisions further along in the spatial modeling process. For example, in the future we might limit the inputs to a maximum likelihood image classification to those with the highest frequency of inclusion in the good models, or a subset of those models. Because the all-subset method yields results similar to those obtained by stepwise selection procedures and provides us with additional information, we consider it valuable for reducing data redundancy and examining multivariate relationships within a modeling dataset.

Literature Cited

DRAPER, N. and H. SMITH. 1981. *Applied regression analysis.* New York: John Wiley and Sons.

LISTER, A.J., R. RIEMANN, J. WESTFALL and M. HOPPUS. *In Press.* Variable selection strategies for small-area estimation using FIA plots and remotely sensed data. In press; to be published in Proceedings of the 2002 FIA Symposium/Annual Meeting of the Southern Mensurationists, November 19-21, 2002, New Orleans, LA.

McGARIGAL, K., S. CUSHMAN and S. STAFFORD. 2000. *Multivariate statistics for wildlife and ecology research.* New York: Springer.

MONTGOMERY, D.C. and E.A. Peck. 1982. *Introduction to linear regression analysis.* New York: John Wiley and Sons.

SAS INSTITUTE. 1999. *User's guide, version 8.* Cary, NC: SAS Institute.

SHTATLAND, E.S., E. CAIN and M.B. BARTON. 2001. The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System, Paper 222-26. In: *SUGI'26 Proceedings.* Cary, NC: SAS Institute Inc.

ABOUT THE AUTHORS

Andrew Lister, Mike Hoppus, and Rachel Riemann are Research Foresters with the USDA Forest Service, Forest Inventory and Analysis, Newtown Square, PA.

PROCEEDINGS

SOCIETY OF AMERICAN FORESTERS
2003 NATIONAL CONVENTION

Forest Science in Practice

Society of American Foresters
5400 Grosvenor Lane
Bethesda, MD 20814-2198
www.safnet.org
SAF Publication 04-01
ISBN 0-939970-86-4