VALIDATION ALTERNATIVES FOR CLASSIFIED IMAGERY

Ben McMahan and Keith T. Weber
GIS Training and Research Center
Idaho State University
Pocatello, ID 83209-8130
(http://giscenter.isu.edu, e-mail: giscenter@isu.edu)

INTRODUCTION

The only reliable means to determine the accuracy of models developed from remotely sensed imagery or geographic information system (GIS) analyses is to perform a validation test. This is typically accomplished by assembling a standard contingency table (or confusion matrix). The matrix is developed by selecting numerous sample points representing each category of the model and determining if the observed field value at that location agrees with the value predicted by the GIS model (e.g., classified imagery). Errors are then reported as omission, commission, and overall error. The Kappa statistic (Titus et al. 1984) gives yet another calculation for classification accuracy, expressed as how much better (or worse) the classification is relative to chance alone.

Validation is typically performed using sample points collected independent of training site data. While this process may seem statistically rigorous, it can be unproductive if researchers have limited field availability or a short field season. In such cases, models will be built in year 1 and not validated until year 2. Changes that can occur in one year can be fairly substantial and it is felt that errors in validation may result because of this temporal delay. To address this, we collected all field data necessary for both model production and validation during the summer of 2002. We then experimented with various bootstrapping (*n* random subsets of the data are

created and tested for consistency of performance) validation iterations to determine the optimum number of field samples needed for classification relative to the number of validation samples.

Within the literature on validation procedures, a number of different methods can be employed to assess the accuracy of the models. One of the most common methods is subset validation (mentioned above), which involves the use of training and test subsets of the data, where the model is built with the training subset and subsequently validated with the test subset. Other more comprehensive validation procedures such as cross-validation (also known as leave-one-out validation) or bootstrapping also exist (Weiss and Kulikowski 1991, Efron and Tibshirani 1993.

METHODS

We iteratively classified fuel load models (with five fuel load categories) using various randomly selected subsets of field data (n = 370) and maximum likelihood classifier. Training site sample sizes were 22, 44, 66, 88, 133, 185, and 370. Validation was performed using all un-used sample points save for the last iteration where all points were used for both model development and validation.

RESULTS

The resulting accuracy of models created with a training subset (n = 185) and validated with a test subset (n = 185) were nearly identical to the results from models developed and validated using the entire dataset (n = 370) (Table 1). During each classification iteration the variance for each spectral signature was calculated for each band of imagery used (i.e., Landsat ETM+ bands 3, 4, and 5) (Figure 1).

Table 1. Comparison of accuracy assessment using 50% of field samples for model developement and 100% for model development and validation.

	All Point Technique	Subset Validation
Overall accuaracy	47.27%	48.28%

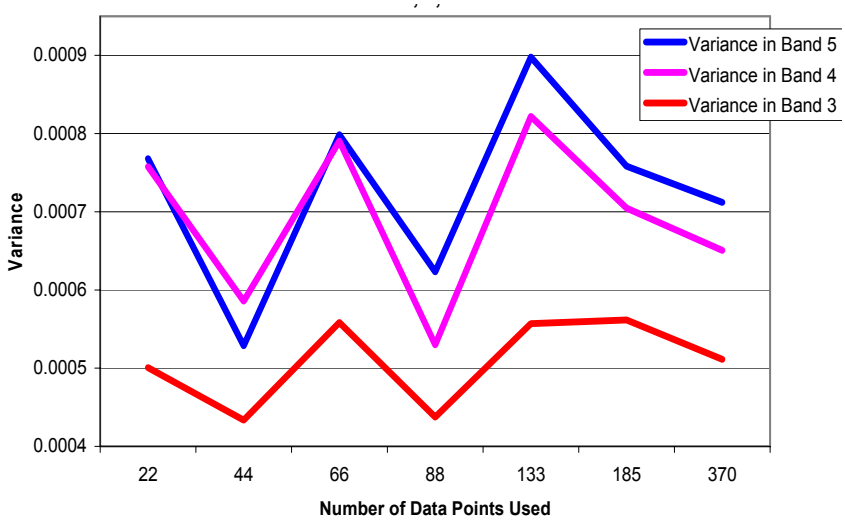

Figure 1. Variance of spectral signatures using different subsets of training sites.

DISCUSSION

The fact that accuracy is nearly identical when using 185 sample points compared to all 370 sample points indicates it is legitimate to perform validation using the same training sites employed to build the model. However, a word of caution is merited. If the total number of points collected is relatively small, then the influence of each individual sample point on the output model (regardless of classification technique selected) will increase. At some point, the classified value at each sample point will perfectly reflect the original value of the sample point. In such cases, accuracy will be 100% and the alert user should suspect a problem. A less obvious example exists as one approaches the scenario described. In this case we expect accuracies to artificially soar. Therefore we consider it prudent to first test the reliability of your sample dataset by performing a 50/50 development/validation classification and compare the resulting accuracy with a development/validation classification where all sample points are used. To employ this technique then, users must ensure they have a collected a large number of field observations that truly captures the variability between and among each category being classified.

ACKNOWLEDGEMENTS

This study was made possible by a grant from the National Aeronautics and Space Administration Goddard Space Flight Center. ISU would also like to acknowledge the Idaho Delegation for their assistance in obtaining this grant.

LITERATURE CITED

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap, London: Chapman & Hall.

Titus, K., Mosher, J.A., Williams, B.K. (1984) Chance-corrected classification for use in discriminant analysis: ecological applications. *American Midland Naturalist*, 111, 1-7.

Weiss, S.M. and Kulikowski, C.A. (1991), Computer Systems That Learn, Morgan Kaufmann.