# P. 7.7 FACTORS AFFECTING VEGETATION COVER MAPPING FOR LANDFIRE

James E. Vogelmann*, Chengquang Huang, Brian Tolk
SAIC, U.S. Geological Survey,
EROS Data Center

Zhiliang Zhu
U.S. Geological Survey,
EROS Data Center

## ABSTRACT

Three dates of Landsat 7 Enhanced Thematic Mapper Plus (ETM+) satellite data, digital elevation model data, potential vegetation information, biophysical settings information, and field data are being used to generate the vegetation map information for a LANDFIRE pilot project in Utah. Extensive field information (6188 plot locations) has been made available to this project from the US Forest Service Forest Inventory and Assessment (FIA) and Utah State University. Digital information from each data layer is extracted for each plot location, and decision tree analysis is being used to generate the land cover classifications. The relative importance of the various data layers for vegetation mapping and determination of the number of field plots necessary for optimal mapping are under investigation. First-order accuracy values are being determined using cross-validation approaches. Research thus far indicates that some vegetation communities, such as pinion pine-juniper, are very distinct and are relatively straightforward to map. However, others, such as Douglas-fir and white fir, are spectrally very similar and represent potential classification challenges. We are working to optimize the use of the available field information with the spatial data in order to generate the best large region land cover products possible in an operational framework.

## 1. INTRODUCTION

The LANDFIRE project is a joint effort between USDA Forest Service and Department of the Interior agencies to provide the spatial data and predictive models required for characterizing fuel conditions and fire regimes and for helping to evaluate fire hazard status. A significant component of the project involves development of a detailed land cover classification data layer that can be used in conjunction with other spatial data layers for input to various fire fuels and fire characterization models. In the current investigation we are conducting a pilot project in central Utah to develop and fine-tune land cover generation methodology to meet LANDFIRE requirements.

There are currently many questions and debates regarding the most practical and efficient ways to generate large region land cover data sets at sufficient levels of thematic detail and accuracy for pertinent modeling applications. We are in the process of determining appropriate techniques and assessing the role of various spatial data layers for generating regional vegetation classification data products appropriate to LANDFIRE. As part of this effort, we are investigating the role of available field information for large area land cover mapping. This includes information collected by Forest Inventory and Assessment (FIA) research units and data collected by Utah State University personnel. Both sources of field information contain a large breadth of plot-based data characterizing the region's land cover, and investigations are being made to determine optimal methodology for incorporating this information into the classification process.

## 2. SPATIAL DATA

Landsat Enhanced Thematic Mapper Plus (ETM+) data are the primary source of remotely sensed information used in the LANDFIRE project. During the first year of the project, we are concentrating on a nine-scene region in the Wasatch and Uinta Mountains of Utah. Three dates of imagery are acquired for each World Reference System 2 (WRS-2) path row, and represent spring, summer and fall time frames.

*Corresponding author address: James E. Vogelmann, SAIC, USGS EROS Data Center, Sioux Falls, SD 57198; e-mail: vogel@usgs.gov

Digital elevation model (DEM) data and derivatives (slope, aspect, position index) are also being used in this investigation. The source of DEM is the National Elevation Dataset (Gesch et al., 2002). Resolution of the DEM is 30 meters. In addition, a series of biophysical data layers generated through modeling of climate, DEM, soils, field data, and other sources of information are being used. These include leaf area index, soil temperature, actual and potential evapotranspiration, and degree-days, as well as many other parameters. An example of a biophysical data layer is shown in Figure 1. Potential vegetation type data layers are also being used in the analysis (e.g. Figure 2). The latter type of information provides spatial depictions of where particular types of vegetation can and cannot exist, and thus should be useful for eliminating certain distribution-related errors.

## 3. FIELD DATA

Field data were collected and provided by personnel from FIA and Utah State University. The FIA Program collects, analyzes, and reports information on the status and trends of forests within the United States. The FIA established a series of permanent 1-acre (0.4 hectare) plots across the United States; forest measurements are now made on one tenth of the sites each year. Included within the FIA data set is information on many standard forest parameters, including species dominance and co-dominance, basal area, and tree height. For the pilot area, plot data were available from 2052 forest sites. Utah State University field data were based on visual assessment (variable plot size), and included vegetation composition information collected for both the current LANDFIRE project as well as for the Utah State University Gap Analysis Program. Data from 4136 plots (2209 forested and 1927 non-forested) were available for the entire pilot area from Utah State University.

## 4. DATA ANALYSIS

Digital values were extracted from imagery and ancillary spatial data layers for each field data plot. Decision tree analysis using the C5 program (Quinlan, 1993) was done using various combinations of Landsat and DEM data sets. One of the advanced features of this program is

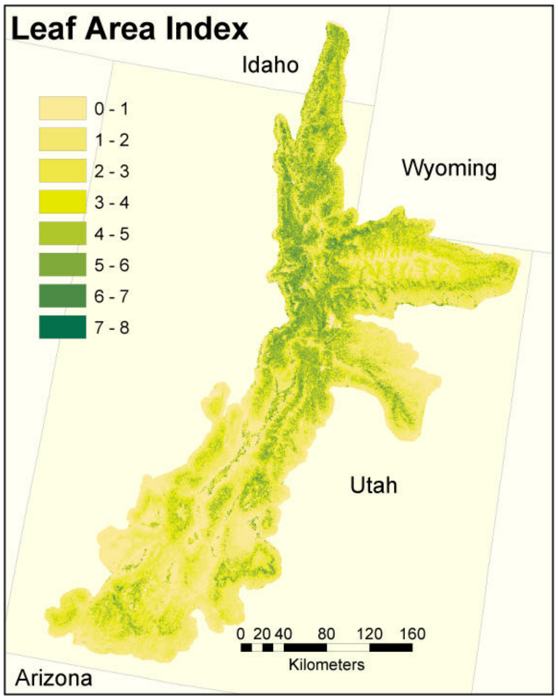

Figure 1. Example of biophysical data layer (leaf area index) for study area.

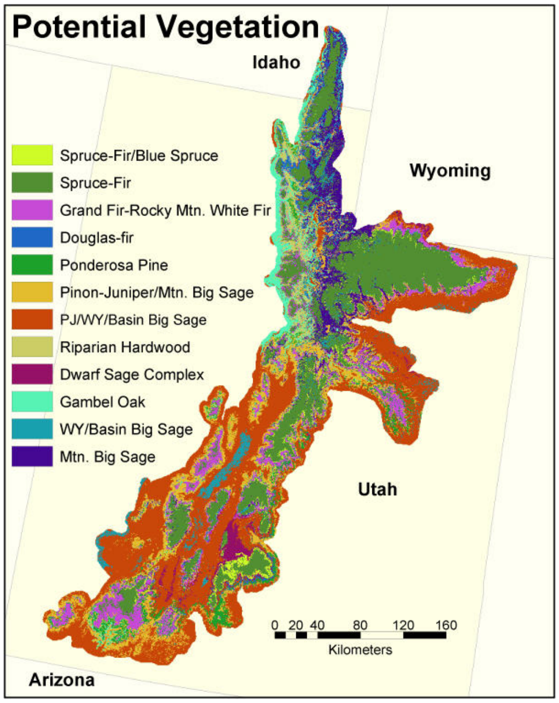

Figure 2. Potential vegetation map of study area.

boosting, a technique for improving classification accuracy (Bauer and Kohavi, 1998). With this function, the program develops a sequence of decision trees, with each subsequent one trying to "fix" the misclassification errors in the previous tree. Each decision tree makes a prediction, and the final prediction is a weighted vote of the predictions of all trees. The program also enables cross-validation, which consists of repeated experiments in which a subset of the sample is used to train a classification model, and an unseen subset is used to evaluate the model. For model runs in this study, the original data sets were divided into 5 equal sized subsets for a 5-fold cross-validation, with each subset being used to evaluate the algorithm trained using the remaining 4 subsets. Cross-validation is ideal for providing first-order estimates regarding the probable accuracies of the classification products that will result from a given decision tree model run.

## 5. RESULTS AND DISCUSSION

A 28-class vegetation type map is shown in Figure 3. Results depicted were generated using decision tree model runs included just imagery, DEM and field information (no biophysical data layers or potential vegetation information). The overall distribution pattern of major vegetation types appears to be reasonable, with spruce/fir dominating the higher elevations, and pinyon/juniper dominating the foothills in the southern part of the study area. Most of the shrub types, such as the different sagebrush communities, are located at lower elevations. This classification had an overall cross-validation accuracy of approximately 60% (stratified by land form).

Model runs of forest type using imagery, DEM data, five biophysical data layers (degree day, actual evapotranspiration, potential evapotranspiration, soil temperature, leaf area index) and potential vegetation information improved results slightly; cross-validation accuracies were approximately 62% for ten forest types. Examination of the cross-validation error matrix indicates that a substantial amount of "error" can be attributed to many pinyon pine-juniper plots being classified as juniper plots, and vice-versa. While technically an "error", it should be noted that in reality these two classes represent a continuum in nature, and that the delineation between the two forest types for LANDFIRE is quite arbitrary. Thus, it is not surprising that there is some confusion between the two classes. If pinyon pine-juniper and juniper are considered one class, overall forest class accuracy improves to 71%. Douglas-fir and white fir had relatively low accuracies, and it appears that these forest classes will be the most difficult to map (accuracy values of 42% and 52%, respectively).

We are in the process of assessing the value of biophysical data layers and potential vegetation information for shrub and grass classes. In addition, we are testing the effects of sample size and plot heterogeneity on model results.

## 6. REFERENCES

Bauer, E., and Kohavi, R. (1998). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning* 5: 1-38

Gesch, D., Oimoen, M., Greenlee, S, Nelson, C., Steuck, M., and Tyler, D. (2002). The National Elevation Dataset. *Photogrammetric Engineering and Remote Sensing*, 68(1): 5-12.

Quinlan, J.R. (1993). *C4.5 programs for machine learning: The Morgan Kaufmann Series in Machine Learning*. San Mateo, California, Morgan Kaufmann Publishers, 302 p.

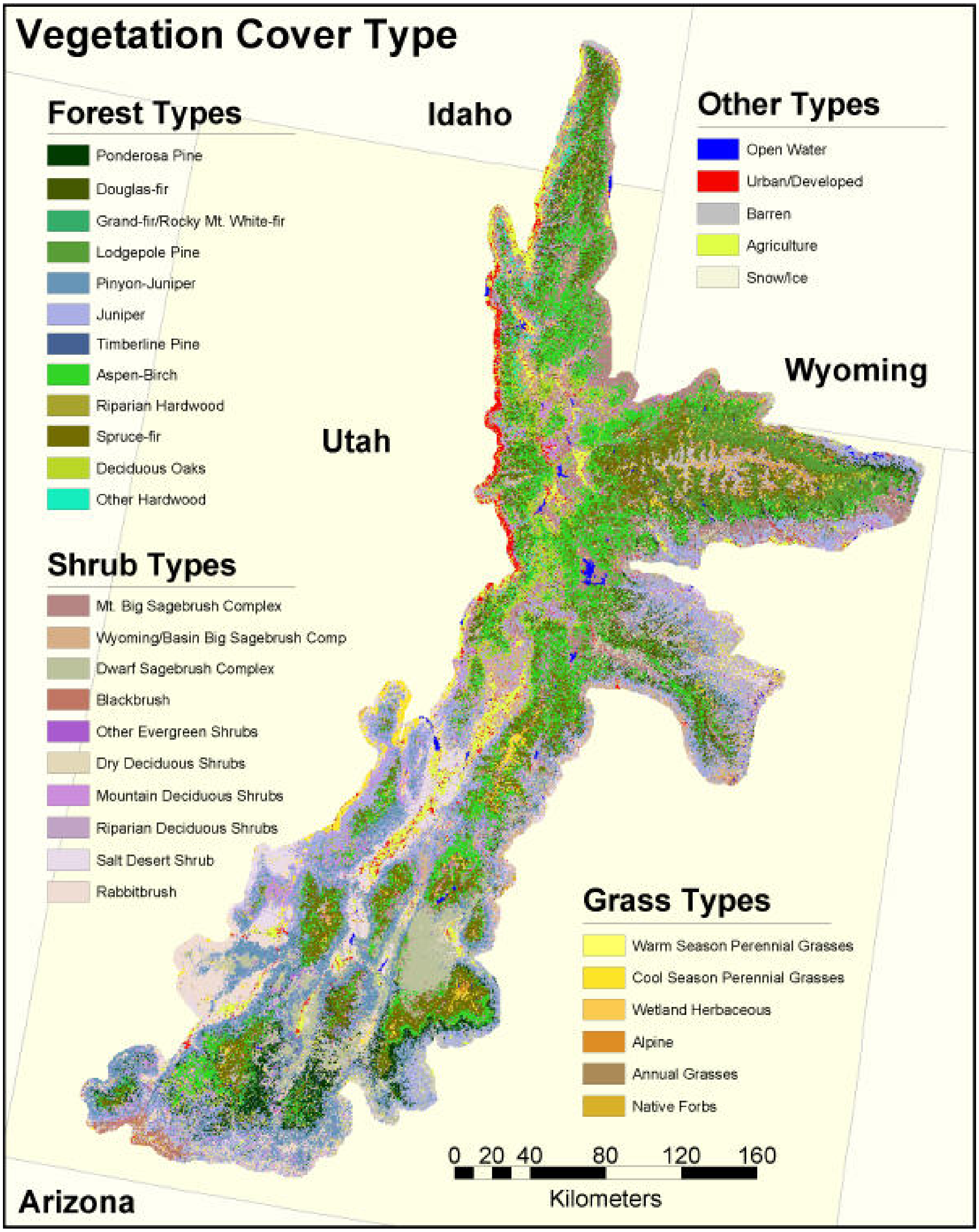

Figure 3. Preliminary land cover type classification for Utah test area.