Testing an Approach to Improving Fire Fuel Mapping by Modeling Fuel Structure and Types Based on Combined Satellite Imagery and Field Data

— Final Report —

Zhiliang Zhu
Research Physical Scientist, USGS/EROS
and
Michael D. Fleming
Senior Scientist, SAIC/TSSC at
USGS/Alaska Science Center

Review of Project Objectives

When this project was proposed, there were no good mapping tools to relate information collected on field inventory plots with remotely sensed imagery, a technique that was needed in order to produce useful wildland fuel data. The project was envisioned to develop and test an algorithm that could accomplish the mapping of key vegetation parameters to meet the needs of the fire science community in test areas of the conterminous U.S. and Alaska. Specific objectives of the project were: 1) Develop a simple, practical methodology (the k nearest neighbor or k-NN) to integrate spatial data from field sample sites and satellite image data. 2) Map several key fire fuel layers including vegetation type, canopy density, canopy height, basal area, and green biomass. 3) Calibrate and validate the data sets, and conduct the technology transfer.

Since the funding of the project, a new national project called LANDFIRE came into existence to map vegetation and wildland fuels at the national scale and ground resolution of 30m. The LANDFIRE project became a natural test case and linkage for the Joint Fire Science funded project. As the result, much of the effort is closely linked to development and execution of the LANDFIRE methodology.

Project Approach

1. Study areas.

Five study areas were eventually selected based on field data and satellite imagery availability and the desire to test the approach in areas of different vegetation characteristics. The four mapping areas are Chesapeake Bay, northern Utah, southern Utah, and the Denali National Park in Alaska. Because of differences in vegetation cover

and availability of field reference data, there was a certain amount of differences in terms of input and output data layers for the study areas.

2. Input predictor data used for the study.

Significant amount of data preparation work was conducted to generate a set of 30m resolution spatial datasets for use in testing the k-NN algorithm and developing resulting deliverables. For all four test areas, a common set of data were used as listed in the table below.

Characteristics of input data used in the study.

Data Type	**Dataset**
Landsat 7 spectral bands (micrometers)	Band 1 (0.45–0.52)
	Band 2 (0.52–0.60)
	Band 3 (0.63–0.69)
	Band 4 (0.76–0.90)
	Band 5 (1.55–1.75)
	Band 7 (2.08–2.35)
Spectral transformations	Tasseled Cap bands 1-3
	Normalized difference of vegetation index
Biophysical gradients	Solar Illumination
	Elevation
	Slope
	Transformed Aspect
	Topographic Position Index
	Potential Total Solar Radiation
	Distance to Streams
	Distance to Major Streams
	Time since last fire (yrs)
	Soil - Average Water Content
	Soil - Carbon
	Soil - Quality

3. Development of a k-NN software package.

A software package implementing the k-NN algorithm needed to be developed in order to test the effectiveness of the algorithm for mapping the vegetation variables. For this study, we used IDL (Interactive Data Language) to develop

the k-NN software. Descriptions and the source code of this software are included in this report.

4. Produce output data layers in an iterative process. The desired output data layers were mapped using both the k-NN and decision tree or regression tree techniques, as well as the artificial neural networks. Various controls and parameters for the k-NN software were tested in mapping the vegetation variables. The table below lists various maps produced for the four study areas.

Output data types produced for the four study areas

Dataset	**Chesapeake Bay**	**Northern Utah**	**Southern Utah**	**Denali National Park**
Forest, non-forest	X	X	X	-
Conifer, mixed, deciduous	X	-	-	-
Forest types	X	X	X	X
Crown cover (%)	X	X	X	-
Total basal area	X	X	X	X
Conifer basal area	X	X	X	-
Deciduous basal area	X	X	X	-
Sapling basal area	X	X	X	-
Average dominant height	X	X	X	X
Total gross biomass, trees, Dry	X	X	X	-
Total gross biomass, trees, green	X	-	-	-
Total gross biomass, sapling, dry	X	-	-	-
Total tree volume, board feet	X	-	-	-
Ground land use	-	X	X	-
Shrub types	-	X	X	-
Grassland types	-	X	X	-

Stand size	X	-	-	-

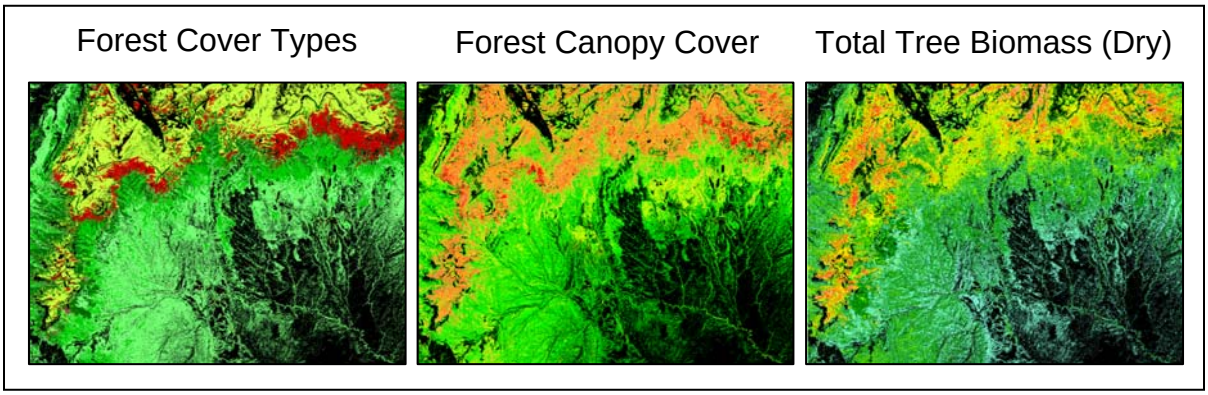
Many results were produced by the project. We evaluated various types of variables that could be classified, and generated meaningful classifications for a large number of types of variables. Results of the four study areas may be summarized succinctly as follows:

- Chesapeake Bay area, majority of parameter study was done on this site, largest number of classification was generated for various forest type and structure variables.

- Northern Utah and Southern Utah. Pair of single full ETM+ scenes located in the northern and southern parts of the study areas. Here we concentrated on mapping the LANDFIRE variables (the vegetation types, tree height and canopy closure vegetation) and continuation of the parameter study.

- Denali, Alaska. We mapped LANDFIRE forest type and structure classes for boreal forest region of Alaska. Successfully separated conifer and deciduous species.

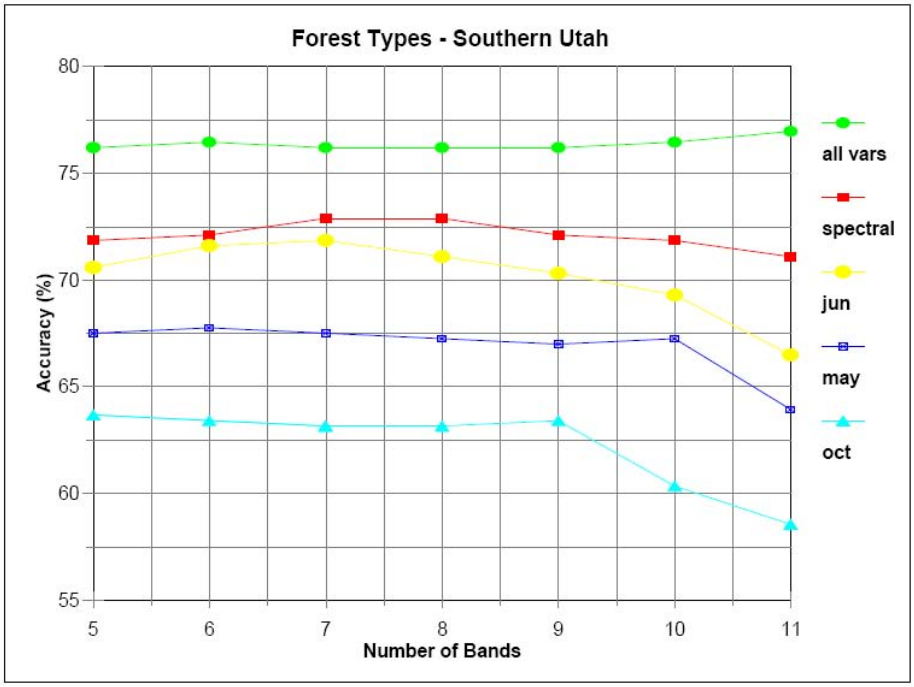
Various vegetation map products were generated for the four study areas using the k-NN algorithm and decision tree and regress tree classifiers. The figure below illustrates vegetation maps produced for the southern Utah study area using the k-NN classifier.



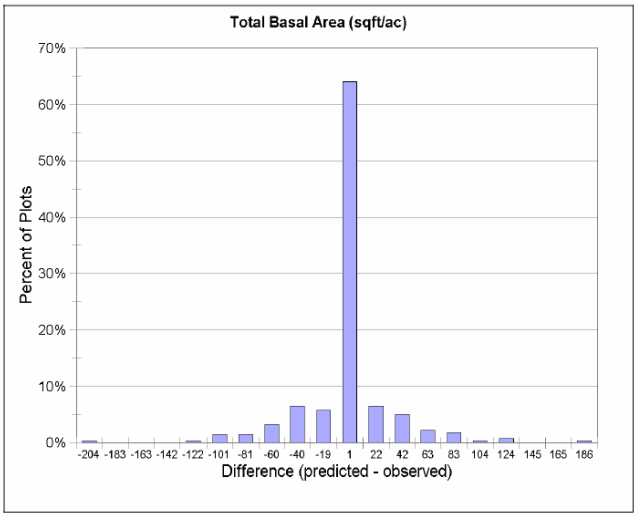
5. Evaluation and validation process

Extensive evaluation and validation of vegetation map products were conducted to derive information on relative performances of k-NN as compared to other methods. Overall mapping accuracy was used as an evaluation tool to measure effectiveness of the mapping. The figure below illustrates mapping accuracy for

forest types in southern Utah as a function of number of spectral bands and dates of Landsat imagery. Mapping accuracies were generally obtained at a range of 70-80% for cover types.

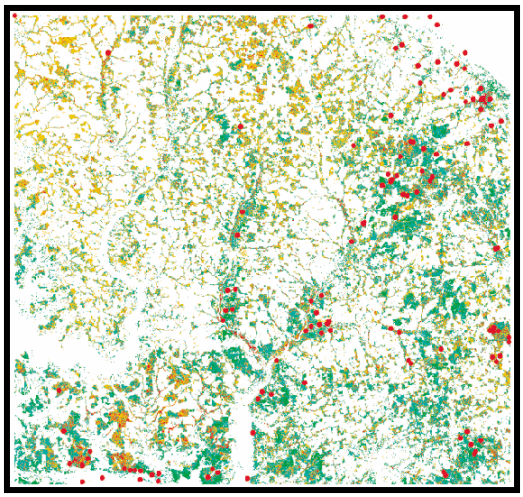


Mapping accuracy was not the only tool used to evaluate effects of the mapping classifiers. We also relied on other quantitative evaluation methods such as the distance image which essentially shows how well the pixel in the image was represented in the field, mean differences and RMSE (Room Mean Square of Error), as well as examination of statistical difference between the mapped classes and the field plots, as shown in the figure below, and indicates that the significant majority of mapped pixels had very little difference when compared to field reference data. Where there was a very large difference between the predicted and the field data, the field data was usually inaccurate; either because of locational errors or changes since the field plot data was collected.



Other evaluation methods included a qualitative evaluation of each of the classifications as they were generated. A visual comparison with an enhanced CIR image of the ETM+ data gave a very good idea how good the classification was and the types of errors. We also used field reference plots to

derive accuracy and correlation numbers for various mapping deliverables. The figure and table below show 1) overlay of field reference plots on forest cover types mapped for the Chesapeake Bay study area, and 2) corresponding accuracy for forest cover types mapped.



Forest-NonForest	Plots	Accuracy
non forest	171	95.3
forest	107	99.1
Total	**278**	**96.76**

Forest Types	Plots	Accuracy
non-forest	171	94.7
loblolly-shortleaf pine	17	58.8
oak-pine	47	70.2
oak-hickory	33	54.5
oak-gum-cypress	9	22.2
elm-ash-cottonwood	1	0.0
Total	**278**	**80.94**

k-NN Software Evaluation

1. k-NN Software Development

   A. Classifier - written in IDL (Interactive Data Language); runtime license available on UNIX, Macintosh and Microsoft Windows platforms; efficient and powerful, array-oriented language with numerous mathematical analysis and graphical display algorithms built in with the following features:

   - Discrete mode to classify categorical variables (classes)
   - Continuous mode to classify quantitative variables (numbers)
   - Output products include a distance image that quantitatively evaluates the classification; the distance is to the k^{th} plot (the longest distance of the plots selected to make the classification).
   - Masking capability to limit the classification to only certain areas in the image.

   B. Evaluation Tool - accuracy estimation using a leave one out cross-validation using the training plot data set(s).

   - Generates error matrices in discrete mode.

- Generated an error histogram, mean difference, and RMSE estimates in continuous mode.

- Used to estimate the potential accuracy of a classification using the training data set.

- An independent test data set can be run using the training data set.

- Used to evaluate band combinations (Feature Selection Tool)

C. Feature Selection Tool - Uses the Evaluation Tool to estimate the accuracy for a number of band combinations.

- Algorithm selects the best subset of bands by running all combinations of bands to identify which subset of available input variables will yield the highest accuracy.

- Can be used to estimate the best band combination while varying any of the algorithm's parameters.

- Evaluated using an indexing system to improve the efficiency of the feature selection and classification algorithms. Initial results were not promising and more study would be needed. Evaluated indexing using single bands, NDVI, and several other of the input variables that were available.

D. Data Extraction Tool - extracts the data values from the input variables at the field data site locations. Used to generate the training plot data sets that are input to the evaluation and feature selections tools, and the classifier.

2. k-NN Classifier parameter study. Evaluated a number of the parameters involved in the analysis using the k-NN classifier in outputting both continuous and discrete variables.

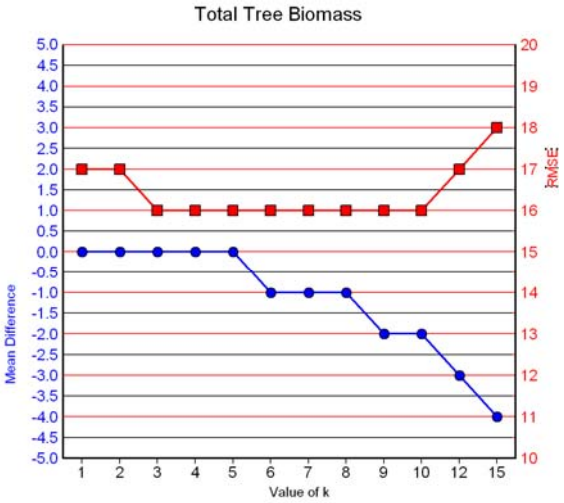
A. The best number of bands to use in the classification was evaluated. As a rule of thumb, the best number of bands is about the same as the number of output classes desired. Less than three bands usually results in a poor classification (too many ties). Not significant effect on the analysis and classification times.

B. Size of sample probably the most important toward getting good classifications, a larger sample is better, to a point. Need to represent full range of variation in each class and only those classes (not so large an area that the classifier becomes confused). Errors in training data caused serious problems in the classification. As the result, the classifier did not

extrapolate well. Large samples require considerable CPU resources, more so for the feature selection analysis. Sample sizes of up to 400 or 500 field plots keep the analysis and classification times manageable.

C. Value of k, the number of neighbors to use to make the classification is important, but not as much as the sample size (n). Values of k = 1 yields a speckled classification with the most detail. As the figure below shows, higher values of k result in a smoother and more general classifications, to the point of eliminating the classes with small sample sizes, because k is larger than the number of plots in the training class.

D. Types of data utilized in the classification were evaluated, spectral only, spectral & topographic Summer TM is the best, multiple dates improves classification, addition of DEM and derived variables results in another jump in accuracy.



E. Combination of bands utilized in the classification was evaluated, which variables were most commonly used and which were not. Band combination selected very important, depends on k, variables available, and number of classes. Soils variables (Z60) were not used, caused artificial lines in the classifications

3. Comparison: k-NN classifier vs. Decision Tree & Regression Tree classifiers

A. Vegetation Type classifications. Forest, shrubs, and grass types were mapped using both the decision tree and k-NN classifiers using exact same training data sets and input images for the entire mapping zone 16. k-NN error matrixes and decision tree accuracies have been discussed previously. General summary of comparison is as follows:

- k-NN generally a little more accurate, but feature selection can require considerable CPU resources for training the classifier.

- Regression and decision tree methods produced similar results, but minimal requirements for CPU resources, although multiple analysis and classifications may need to be generated to get a good set of

bands selected for the classifier, but not as sensitive to which bands are input to the algorithm.

- k-NN classified better for simpler problems, the regression and decision tree was better in more complex problems.

- k-NN in continuous mode worked a better that the equivalent regression tree classification.

- Sample size is the major limitation, some training classes had few plots and some vegetation types were not sampled at all (riparian willow/cottonwood community in southern Utah).

Review of Project Deliverables

Major deliverables of the JFSP funded project include the following:

- Development and testing of a k-NN software for mapping vegetation structure and composition

- Produce maps and associated accuracy numbers of the vegetation structure and composition over test areas

- Produce technical papers and conduct technical presentations

- Conduct technical transfer; ensure that the results were used for operational mapping of wildland fuel parameters.

Project Accomplishments (included in separate folders)

1. k-NN Software Development: (source code included with the report).

   - Classifier - with both discrete and continuous modes, and a quantitative evaluation image.
   - Evaluation Tool - accuracy estimation using a leave one out cross-validation that generates error matrices or an error histogram, mean difference, and RMSE estimates.
   - Feature Selection Tool - used to select the “best” band combination, a subset of all available input variables (imagery, topographic, distance2x, ...).
   - Data Extraction Tool - used to extract the data values from the input variables at the field data site locations.

2. Produce vegetation maps that could be used for mapping fire fuels

- A number of map produced included with the report.
- Vegetation parameters mapped include vegetation composition/cover types, canopy cover, canopy height, basal area, and aboveground biomass.

3. Produce technical papers, posters, and presentations

   - Xian et al., PECORA 15, 2002.
   - Zhu et al., ASPRS, 2003
   - Fleming et al., the 2nd International Fire Ecology Conference, 2003
   - Fleming et al, the USGS Fire Science Workshop, 2002
   - Various technical presentations

4. Technology Transfer

   - Number of workshops: Alaska Fire community, Alaska AGDC Land Cover Mapping Sub-Committee, and USFS FIA Regional office (Newtown, PA) to present techniques for mapping vegetation and wildland fuel parameters.
   - The k-NN software has been made available to the LANDFIRE project for use in mapping vegetation composition and structure. Thus, ultimately, the technology is being used by the target user community.

Major Findings of the Project

The mapping of existing vegetation is a core requirement of meeting the fire management community's needs for wildland fuel data. If supported with an adequate amount of field reference data, target minimum accuracies of 60 percent or better are achievable for a mid-level cover type classification at the regional scale, using either the k-NN classifier or the decision tree method. The relation between mapping accuracy and a mid-level classification (for example, 30 vegetation cover types or more in a regionally-sized map area) is highly elastic depending on how the classification is designed. The addition or subtraction of a floristically or ecologically similar cover type could have significant effects on resulting accuracies. Of the three major life forms, herbaceous cover types are the most difficult to map because these species adapt to many general biophysical characteristics and have few unique spectral signatures.

In this study, vegetation structure is defined by canopy cover, canopy height, forest basal area, and above ground biomass. These structure attributes can be

modeled consistently and the k-NN software offers the option of mapping these attributes as continuous variables.

The k-NN classifier performed similarly to that of the decision tree and regression tree techniques. k-NN performed better for relatively simple classifications and for herbaceous and shrub land cover. On the other hand, decision and regression tree classifiers was computation intensive and required hours of CPU time to do the feature selection process, a process that regression and decision tree classifiers took only a few minutes. Optimization of the k-NN procedure became a severe limitation issue.

The incorporation of selected biophysical gradient layers in the mapping models contributes to an increase in mapping accuracy. In addition, the use of the biophysical and ecological stratifications that describe the environmental effects on species establishment and growth also contributes to enhanced mapping consistency.

Acknowledgement

We acknowledge the funding provided by the Joint Fire Science Program. The cooperation of the USFS FIA offices in Newtown Square, PA, Ogden, Utah and Anchorage, Alaska are gratefully acknowledged. The research was performed in part by the Science Application International Corporation under U.S. Geological Survey Contract 1434-CR-97-40274 and 03CRCN0001.