Corresponding Author

John S Hogland
Rocky Mountain Research Station, USDA Forest Service
200 E. Broadway, Missoula, MT 59807
jshogland@fs.fed.us

# Estimating forest characteristics using NAIP imagery and ArcObjects

John S. Hogland[1] Nathaniel M. Anderson[1], Woodam Chung[2], and Lucas Wells[2]

[1] US Forest Service, Rocky Mountain Research Station, Missoula, MT
[2] University of Montana, College of Forestry and Conservation, Missoula, MT

## *Abstract*

Detailed, accurate, efficient, and inexpensive methods of estimating basal area, trees, and aboveground biomass per acre across broad extents are needed to effectively manage forests. In this study we present such a methodology using readily available National Agriculture Imagery Program imagery, Forest Inventory Analysis samples, a two stage classification and estimation approach, .Net numeric libraries, Spatial Analyst, Function Datasets, and ArcObjects.

## *Introduction*

Forested public lands of the United States of America (USA) are a mosaic of diverse ecological communities. In a broad context, a great deal of information is known about these lands (Oswalt et al. 2012). However, much of this information is derived in a very coarse manner and lacks the needed precision, accuracy, and resolution to help determine where and when management activities should occur. Due to the lack of fine grained, timely, accurate, and precise information, managers often follow a less then optimal approach to recognizing potential problems and identifying locations where appropriate silvicultural treatments are needed. This problem solving approach can be described in a series of steps that first detects a broad scale problem once it has become a considerable issue (e.g., insect infestation), finds known locations where the problem exists, and then performs some kind of treatment to remediate the problem. While often thought of in the context of remediation, this same problem solving approach is used when looking to enhance forest benefits (e.g., improving habitat for a particular species of concern or interest). Decision making in this manner tends to be reactionary and expensive, often resulting in partial mitigation that provides fewer benefits than would be expected from optimized outcomes in a data-rich decision making environment.

In a data-rich environment, where forest characteristics are accurately quantified at fine spatial and temporal resolutions across large extents, managers could identify problems and opportunities quicker than without those types of data and could better plan and allocate limited financial resources to meet management objectives. However, attempts to estimate these characteristics at fine resolutions using classical inventory and monitoring methodologies are extremely expensive which often force managers to sacrifice resolution for coverage. To better understand the impacts of this tradeoff, it helps to look at the classical forest inventory and monitoring approach.

From a statistical standpoint, inventory and monitoring endeavors make use of a sample drawn from a population. Observations of the sample are aggregated together to describe characteristics of the population. For communities across a landscape this process generally consists of randomly or systematically selecting sample locations within the landscape and measuring community characteristics for multiple small subsets of that landscape (i.e., the plot). Measurements are then summarized and averaged across all plots to produce estimates of mean and variation for that landscape (Avery and Burkhart 1994). One common extension of this approach uses stratification to partition the landscape into similar areas that theoretically have less variation than the landscape as a whole. Samples are then allocated randomly within each stratum, measured, and aggregated at the landscape level using a weighting procedure based on the proportion of landscape area in each stratum (Avery and Burkhart 1994).

Often the estimates generated from these types of samples are scaled to some generally accepted unit of area (e.g., acre or hectare) and expanded based on the total area of a subset of the landscape (e.g., polygons) to produce an absolute estimate for each subset. In the context of all polygons within a measured landscape, individual polygons have the same interpretation, however interpreting polygon level estimates may have little meaning in a project planning context and in many cases may actually be misleading. Simply stated, mean and variation estimates of forest community characteristics calculated in this manner can differ substantially from one another at landscape and project levels.

To estimate mean and variation of forest community characteristics at the project level, a more intensive sample scheme is implemented. In this case, the population is defined by the area within the geometric boundary of the project (i.e., the total area of the project, such as a forest stand). Navigating and measuring community characteristics at each plot occurs in the same manner as at the landscape scale, however mean and variation estimates of community characteristics now relate to the project as opposed to the landscape as a whole. For example, in forestry a silvicultural prescription for a timber harvest is typically based on a systematic sample of variable radius plots collected along a grid in a stand, at a plot intensity that meets some threshold for estimating standard error. These estimates are then used to evaluate and compare multiple treatments and their potential outcomes at the finer project scale. Here again scale plays an important role and the potential impacts of using project data to describe the landscape as a whole may be unjustified, depending on factors such as landscape heterogeneity and sampling design. This tradeoff between resolution and coverage introduces a high level of uncertainty to the process of forest management planning and implementation, potentially resulting in undesirable outcomes.

Theoretically, managers could forgo the broad scale landscape sampling scheme by defining contiguous project areas across the entire landscape, sampling and estimating the community characteristics for each project area, and finally aggregating each project estimate to produce a landscape level estimate. However the high cost associated with this intensive sampling is typically prohibitive. Moreover, it is inefficient if managers only treat a small portion of the landscape each year. Managers may not know the timing, scope and boundary of a project much before its creation, and project areas that were previously well defined may be split or aggregated into new areas depending on current conditions on the ground.

As an alternative to the classical way community characteristics are estimated (i.e. aggregation across geometric space), community characteristics could be estimated by developing relationships with other attributes, such as spectral reflectance of imagery, at the plot level for a predefined landscape. Estimates derived in this manner extend the classical spatial aggregation technique to include variables that have significant correlation with community characteristics, thereby reducing variation in mean estimates (i.e., resulting in more precise estimates) at potentially very fine spatial and temporal resolutions. For example, in the case of using imagery from the National Agricultural Imagery Program (NIAP), data spatial and temporal resolutions are 10.76 feet$^2$ (1 meter$^2$) every 5 years. In this scenario, sampling intensity and associated sampling cost relates to the amount of variation between spectral and measured values allowing estimates of forest community characteristics to vary independently from space, across the landscape. Moreover, relationships derived between measured values and spectral values often require fewer samples to produce equivalent estimates of community characteristics when compared to aggregating across space alone. For example *Populus tremuloides* (Quaking Aspen) basal area can vary substantially across a project (geometric space), but may vary little in relationship to the texture of spectral reflectance within that project (feature space).

While this intuitively makes sense, “a picture [imagery] is worth a thousand words”, relatively few inventory and monitoring projects use imagery directly to estimate community characteristics. Instead, imagery is used indirectly to define strata (Scott et al. 2005) and classify community types (Brown & Barber 2012). Strata and community types are then used to partition the landscape into similar areas and samples are aggregated as previously described. Still others use imputation to partition samples into unique portions of multidimensional spectral space. New observations are then allocated sample values or weighted mean values of samples based on the proximity of that observation in spectral space to its closest neighbor(s) (Crookston and Finley 2008, Wilson et al. 2012). Although using imagery in this manner often reduces spatial variation in community characteristics, it comes at the cost of the spectral and spatial fidelity of the imagery. To some degree, these types of aggregation techniques continue to be common in the management community because of tradition, concerns about continuity in analysis, project objectives, lack of readily available imagery, computer processing and storage constraints, and limited statistical and machine learning algorithms within standard remote sensing and geographic information system (GIS) software. While advances in technology have alleviated some of these issues

and have made imagery products from programs such as NAIP and Landsat readily available (USGS, 2012), relatively few organizations have systematically embraced the high potential of these technologies and datasets to deliver accurate and precise estimates of forest characteristics across large landscapes at relatively low cost.

## *Methods*

### *Overview*

To this end, we present a methodology that extends the classical estimation approach by relating field measured values of forest characteristics to summarized textural values of high resolution NAIP derivatives within a defined landscape. This approach requires both response (forest characteristics) and explanatory variables (texture derivatives) to be spatially located. U.S. Forest Service Forest Inventory and Analysis Program (FIA) data (US Department of Agriculture Forest Service 2012) and NAIP color infrared (CIR) imagery (Mauck et al. 2009) are two such datasets. While response variables may represent a subset of the population (i.e., sample), there must be a complete inventory of explanatory variables across the landscape. Using the spatial locations of the response variable values, explanatory variable values are extracted and related to the response to create a predictive model. The resulting model can then be applied to the full coverage inventory of explanatory values to produce a prediction surface of the response across the landscape, along with associated estimates of error.

To illustrate this methodology we present a case study in the Uncompahgre National Forest (UNF) that draws direct relationships between forest community characteristics (FIA data) and texture derived from high resolution imagery (NAIP). The boundary of the UNF contains approximately half a million acres of forest, shrub, and grassland communities (Figure 1). For this study we were interested in estimating basal area per acre (BAA), trees per acre (TPA), and tons of above ground biomass per acre (AGB) within multiple forested communities across the boundary of the National Forest at both landscape and project scales. Sampling intensity consisted of 83 field plots. Each field plot was made up of four subplots: one located at plot center and three each located at 120 feet from plot center at bearings of 0, 120, 240 degrees (Figure 2). Field plot data used to develop BAA, TPA, and AGB models were collected by FIA using the FIA sampling protocol (US Department of Agriculture Forest Service 2012). FIA plot centers were located using a navigation grade global position system (GPS) with a horizontal accuracy of $\pm$ 5 meters. Aerial photography for the UNF was collected, preprocessed, and mosaicked by NAIP at a spatial resolution of 10.76 feet$^2$ and a geometric accuracy of $\pm$ 3 pixels (Mauck et al. 2009). FIA plot measurements were related to visually interpreted patterns of 2009 NAIP color infrared imagery (CIR) using a two stage classification and estimation approach.

The first stage of this approach resulted in a probabilistic classification of visually identifiable patterns within NAIP using polytomous logistic regression (Hogland et al. 2013a), derivatives of NAIP spectral reflectance, and second order Gray Level Co-occurrence Matrix (GLCM; Haralick 1973) texture of those spectral derivatives. Probabilistic outputs from the first stage classification were then summarized for the extent of a FIA field plot (window size of 78 x 70 NAIP pixels) using focal and GLCM metrics and were related to FIA plot measurements and estimates of BAA, TPA, and AGB using multivariate linear regression. AGB plot estimates were derived from tree diameters at breast height (DBH; 4.5 feet) using the equations found in Jenkins et. al. (2003).

### *Coding libraries*

Due to the size of NAIP raster datasets (approximately 28 gigabytes), and the types of analyses performed to estimate BAA, TPA, and AGB for the UNF, we developed a parallel coding project that focused on reducing raster processing time and storage space while integrating a wide variety of statistical and machine learning algorithms directly into ESRI's Desktop application (Hogland and Anderson 2014).

While this project initially started out as a set of coding libraries that extended the functionality of Desktop, with relatively little additional effort a graphical user's interface (GUI) was added and deployed as an ESRI add-in toolbar. Through the GUI, the procedures used in this study and described in this paper are now readily available to geographic information systems (GIS) analysts. While these libraries and tools played an import roll in facilitating the estimation and manipulation of BAA, TPA, and AGB outputs, it is beyond the scope of this study to describe the conceptual framework behind these tools and associated algorithms. A more detailed description of coding libraries and user interface that facilitated this analysis can be found in Hogland et. al. 2013b, Hogland and Anderson 2014, and at our project's website (RMRS 2012).

*Stage 1 probabilistic classification*

Prior to performing the probabilistic classification, a 3 x 3 standard deviation, first order texture, focal analysis was performed on each of the NAIP spectral bands. Texture outputs were rescaled to eight bit pixel depth and combined with each of the NAIP spectral bands (green, red, near infrared) to create a six band raster dataset composite. A principal component analysis (PCA), based on the variance covariance of the composite raster dataset (ESRI 2010a), was performed to reduce the dimensionality of the base data and produce an orthogonal transformation of the spectral and textural values. The top three component's Eigen vectors were used to create a multiband raster dataset (PCA raster dataset) where each raster band pixel values were independent of the other raster bands pixel values. GLCM values were calculated (Haralick 1973, Hogland and Anderson 2014) for a 3 x 3 window of the first and third principle components and were used as potential explanatory variables for the probabilistic classification (Table 1).

A sampling intensity of 2000 was deemed sufficiently large to capture the variability between classes and explanatory variables and was determined using a heuristic rule of 30 samples for each class and variable (Foody et al. 2006). To estimate this sample size we assumed that there were roughly 13 visually identifiable patterns (classes) that would help in estimating forest characteristics and five predictor variables used to differentiate each of those classes. Using these assumptions, our heuristic rule indicated that we needed to collect 1950 samples, which was rounded up to a total sample size of 2000 pixel locations. To help insure sampling across the full range of explanatory variable values, the NAIP imagery was spatially split into twenty categories (strata) using ESRI's iterative self-organizing (ISO) clustering routine (ESRI 2010b). One hundred samples (100 pixels) were randomly selected within each of the 20 strata and visually interpreted as a one of 15 identifiable patterns (Table 2). For each selected pixel, a point feature was created and attributed with the importance (weight) of the sample using strata area and PCA and GLCM predictor values for that pixel's location. Interpreted patterns, sample weights, and predictor values were then imported into Statistical Analysis Software (SAS) version 9.2 to develop a parsimonious polytomous logistic regression (PLR) model (Hogland et al. 2013).

Initially, there were 55 potential explanatory variables that could be used to create alternative probabilistic models. To help reduce the number of explanatory variables used to build our final suite of models, we ran an initial PLR classification that selected significant variables using a stepwise selection procedure (SAS/STAT(R) 2010a). The stepwise selection procedure iteratively adds variables that significantly improved model fit and remove variables that do not significantly increase model fit using defined level of significance (alpha). For this stepwise procedure, variables were allowed to enter the PLR model if they significantly improved model fit at an alpha level of 0.15 and were allowed to stay in the model if they continued to improve model fit at an alpha level of 0.10. From the remaining statistically significant variables, seven potential models, each with different combinations of independent explanatory variables, were evaluated using Akaike's (1973, 1974) information criteria (AIC). Explanatory variables for each of the seven potential models were deemed independent of one another based on their variance inflation factor (VIF; SAS/STAT(R) 2010b). Explanatory variables that had VIF values greater than three were sequentially removed from each candidate model until all model variables VIF values were less than or equal to three. The most parsimonious model was then selected based on the

fewest number of variables and lowest significantly different AIC values (i.e., $\Delta AIC > 2$ between competing models).

Using the top ranking model estimates and explanatory raster surfaces, we created a multiband raster dataset depicting the probability that each pixel belongs to one of the fifteen visually identifiable patterns. Each band of the multiband raster dataset, though technically representing the probability of a visually identifiable pattern, can be viewed more generally as a mathematical transformation of the independent PCA and GLCM texture values. Transforming the data in this manner presents them in a context that is easy to interpret as understandable cover types (e.g., "pine") and makes them easy to combine at the same spatial resolution as the FIA field sample without sacrificing the spectral, spatial, and temporal resolution of the data. While it is often useful to apply labels to these probabilistic surfaces (e.g., Shadow and Aspen canopy in Table 2) to facilitate interpretation, it is important to remember that these labels really represent spatial patterns that are deemed useful and identifiable by the analyst in defining relationships to the characteristics trying to be estimated. This means that these labels may or may not accurately describe the visually identifiable patterns within the imagery and that while the analyst may think that a pattern represents Aspen tree canopy, there has been no attempt to verify that in reality that label truly represents the canopy of Aspen tree species. Instead, the label represents a visually identifiable consistence pattern within the imagery, thought by the analyst to have a strong relationship to plot level measurements of community characteristics like tree species BAA, TPA, and AGB. In other words, the analyst need not know exactly what is being observed on the imagery, only that an identifiable pattern exists.

*Stage 2 forest characteristics estimation*

Once built, probabilistic surfaces were combined to identify edge characteristics between shadows and tree, shrub, grass, and bare ground using logical and arithmetic functions and all surfaces were aggregated using Focal and GLCM procedures for a window size of 78 x 70 pixels (Hogland and Anderson 2014; Table 3). Window size and aggregation methodology were chosen based on the extent of the FIA plot (Figure 2) and the anticipated correlation to BAA, TPA, and AGB of common tree communities within the UNF. Aggregated values for each procedure performed were attributed at the spatial resolution of the NAIP pixel making 310 potential predictive surfaces that could be readily sampled using the central coordinate of a given FIA plot (Table 3).

FIA measurements of BAA and TPA were summarized at the plot level by species and scaled to per acre values. Plot biomass in short tons (t) were calculated for the stump (SAGB), bole (BAGB), top (TAGB), foliage (FAGB), and total AGB using Jenkins et. al. (2003) equations and the RMRS Raster Utility command FIA Summarized Biomass (Hogland and Anderson 2014). FIA field data were obtained from the FIA data mart download (FIA 2014) and plot locations for UNF were acquired from an approved project request of the FIA program (Blackard 2011). Common community values of BAA, TPA, and AGB were calculated by summing the individual species values of each community found in Table 4 at the plot level. Potential explanatory variable values for each FIA plot were extracted and attributed to each FIA plot based on the nearest pixel proximity to the center of the central FIA subplot location.

Similar to the first stage classification methods, a subset of potential explanatory variables related to BAA, TPA, and AGB for each forest community were obtained using stepwise variable selection procedure (SAS/STAT(R) 2010a) within the multivariate linear regression procedure. Variables that statistically improved model fit at an alpha level of 0.15 were allowed to enter the model and were allowed to stay in the model if that parameter continued to statically improved model fit at an alpha level of 0.10. From the subset of statistically significant potential explanatory variables, seven multivariate linear regression models were created (Table 4) using explanatory variables that were thought to have strong relationships with community BAA, TPA and AGB. For each forest community, variables that were highly correlated with one another were identified and removed using VIF and a threshold of 3.0 (SAS/STAT(R) 2010b). Our top fitting most parsimonious model for each community was determined by comparing AIC values ($\Delta AIC$) of AGB models and selecting the model that explained the most

information using the fewest parameters. The AGB forest community characteristic was used to compare models based on the relationship between AGB calculations and tree diameter and TPA. Specifically, AGB for a plot was calculated using both tree diameters and the number of trees found within a given plot and in this sense represented a good overall metric of both BAA and TPA.

Slope and intercept estimates for our top fitting multivariate linear regression models were used in conjunction with corresponding explanatory raster surfaces to produce a multiband community raster dataset depicting mean per acre values of BAA, TPA, AGB, SAGB, BAGB, TAGB, and FAGB for each forest community (Table 4). From our multiband community raster datasets, five random locations were sampled between 164 and 1805 feet from a navigable road, according to FIA protocol, within each of the twenty ISO cluster classes used to stratify the UNF for visual interpretation in stage 1. Actual mean field measurements of community BAA, TPA, AGB, SAGB, BAGB, TAGB, and FAGB were then compared against predicted values using weighted residual analysis.

## *Results*

*Stage 1 probabilistic classification*

The first three principal components explained approximately 98.5 percent of the variation within the NAIP CIR imagery and corresponding first order texture measures (Table 5). The Eigen vectors of these three components can loosely be interpreted as measures (i.e. indices) of NAIP brightness, surrounding spatial variation in NAIP pixel values (texture), and a vegetation index. Using the brightness and vegetative index, we built an additional thirteen GLCM raster datasets for a horizontal and vertical offset of one pixel. In total, we created 55 potential explanatory variables (Table 1) thought to have some relationship to our visually identifiable patterns. From those 55 potential variables we were able to remove 45 variables using the stepwise procedure. Using the remaining 10 potential explanatory variables we created and compared seven probabilistic classification models based on ΔAIC values (Table 6). Our top fitting most parsimonious model used the brightness, texture, and vegetative index from the PCA analysis along with horizontal Homogeneity derived from the vegetation index of the PCA analysis and three GLCM texture parameters derived from the brightness index (horizontal Homogeneity, horizontal Correlation, and vertical Correlation; Figure 3). Evaluating our model fit in terms of information gain, our top model significantly ($X^2_{df=98}$, 4551.68; p-value <0.0001) explained 86 percent of the information within the data (max rescaled $R^2 = 0.86$; SAS, 2011). Moreover, every variable in our classification model significantly improved model fit (Type III Analysis of effects p-value < 0.001). Using this model, we created a multiband raster surface depicting the probability of each visually identifiable pattern across UNF (Figure 4).

*Stage 2 forest characteristics estimation*

The outputs of our PLR classification were then summarized for a focal window of 78 x 70 pixels using Focal and GLCM procedures (Hogland and Anderson 2014; Figure 5) and related to the FIA mean estimates of species BAA, TPA, AGB, SAGB, BAGB, TAGB, and FAGB through multivariate regression (SAS/STAT(R) 2010a). From the potential 310 predictors of forest community characteristics (Table 3), we were able to remove 291, 286, 290, 300,283, 301, and 273 variables from Aspen, Fir, Juniper, Pine, Pinyon, Scrub Oak, and Spruce forest communities, respectively, using the stepwise procedure. For the remaining variables of each forest community, we created seven potential predictive models (Table 7) and compared them using ΔAIC. Our top fitting models (Table 8) for forest communities explained between 13 and 72 percent of the variation in the data across BAA, TPA, and AGB and on average explained 62 percent of the variation within the more common forest communities (Figure 6). In all cases across forested communities, estimates of slope for our texture based explanatory variables were significantly different than zero (Table 8) for at least one the forest characteristics. Using the texture values derived from the probabilistic classification, the estimates of slope and intercept for

each forest community and characteristic model, and our coding libraries (Hogland and Anderson 2014), we created seven multiband raster datasets at a spatial resolution of 10.76 feet$^2$ (Figure 7). Each band of the seven raster datasets corresponded to one of seven community characteristics: mean FIA plot estimates of BAA, TPA, AGB, SAGB, BAGB, TAGB, and FAGB. In some cases our community characteristic models predicted negative values. These values were an artifact of the statistical modeling process and were converted to zero in the final raster datasets to reflect the true ground condition.

Using the spatial coordinates of our independent validation dataset and the corresponding values of our predictive surfaces for those locations, we independently evaluated our forest community characteristics models. Across all estimates of forest community characteristics our validation dataset did not violate the assumptions of linear regression (Figure 8). While some models of forested community characteristics over predict or under predict low and high values of a given community, in general deviations from actual values were relatively small and cancel each other out when summarized across the landscape (Figure 9). This is not to say that for any single prediction (i.e., a pixel) there was not a large difference between observed and predicted values (Figure 9B), but rather this outcome implies that, on average, our predictive models accurately predict mean forest characteristics (Figure 9A, inset). In the context of a forestry related project, which typically contains many raster pixels (e.g., a 50 acre project contains 202,323, NAIP pixels), these outputs can be summarized to produce accurate estimates of mean forest characteristics in a similar fashion to what is depicted in Figure 9.

## *Discussion*

Our two stage classification and estimation approach successfully improved estimates of mean forest characteristics across the UNF when compared to classical estimation approaches. We can infer that our models are an improvement by interpreting model fit statistics calculated from the multivariate linear regression modeling process (e.g., correlation coefficient, ΔAIC, and F-test; Table 8). While our predictive model fits are not perfect (e.g., $R^2 < 1$, RMSE > 0), they explain a substantial amount of the variation within BAA, TPA, AGB, SAGB, BAGB, TAGB and FAGB across the forest community types in a linear and easy to interpret fashion based solely on the reflectance and texture of visually identifiable patterns in NAIP imagery.

The first stage of our modeling process accurately and precisely quantifies those visually identifiable patterns using transformations of the NAIP imagery and its texture. These visually identifiable patterns significantly improved model fit for forest community characteristics over texture derived from PCA values alone, and they provide a more meaningful set of variables for interpretation. For example, the linear model of AGB for Aspen forest communities included explanatory variables that quantify the amount of Shadow and Grass edge (SGA), horizontal GLCM Correlation for brightness component of the PCA transformation (PCOR), vertical GLCM Contrast of Aspen cover (ACON), and vertical GLCM Correlation of Dead Grass (DCOR) for an analysis window of 78 x 70 pixels. The slope estimates for these variables are 0.023, 47.533, 0.189, and -20.489, respectively. By interpreting these slope estimates in the context of the linear regression model we can see that as explanatory values increase, Aspen forest community AGB increases for all but DCOR. These relationships make intuitive sense from a biological perspective. Specifically, as the amount of SGA (a correlate for canopy openings), PCOR (a measure brightness similarity), and ACON (Aspen Crown density) increases, we would expect to see an increase in Aspen community AGB. Similarly, as DCOR (a measure of similarity in Dead Grass probability) increases, we would expect to see a decrease in Aspen Community AGB.

More importantly for forest managers, our predictive models are derived in feature space (texture of visually identifiable patterns) and are only bounded by geometric space (boundary of the study). This means that we can use our models to depict spatial variations in forest characteristics across the landscape at the resolution of each raster cell (Figure 7), which is highly preferable to only being able to predict these values for a project area polygon or for broad vegetative strata. Using our newly developed coding libraries, these depictions, in the form of raster datasets, can be easily manipulated within a GIS to

address a wide range of management related questions both across a large area and at the project level, or even across ownerships in a collaborative land management framework. Furthermore, these types of datasets provide managers with the basic information needed to quickly identify potential forest related problems through visual interpretation or automated analysis, and to plan solutions at both landscape and project scales for a fraction of the cost it would take to obtain similar estimates using conventional techniques. These datasets can also be shared without compromising the confidentiality of plot locations, which is a major commitment of the FIA.

Although we illustrated these techniques for seven forest communities found in the UNF, they are equally applicable to a wide range of ecological systems and resource management questions, and can be deployed using a wide variety of datasets that are spatially located with a high degree of fidelity. Similar models can be built and applied across varying landscapes using field and remotely sensed data. Here again, the outputs of models developed in this way can be used to evaluate model fit, build new predictive surfaces, and address numerous resource management questions all within a GIS.

## *Conclusion*

ESRI's object library provides a unique platform that has been leveraged to streamline and facilitate our two stage classification and estimation approach. Our approach has been used to accurately and precisely depict common forest community characteristics at fine spatial scales across a relatively large landscape. These depictions, in the form of raster datasets, provide managers with the information needed to actively identify potential resource related challenges and plan strategies that more effectively manage natural resources at both landscape and project scales.

## *Acknowledgements*

This project was supported by the Rocky Mountain Research Station and Agriculture and Food Research Initiative, Biomass Research and Development Initiative, Competitive Grant no. 2010-05325 from the USDA National Institute of Food and Agriculture.

## *References*


Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory. Edited by Petrov BN and Csake F. Akademiai Kiado, Budapest, pg. 267-281.

Akaike H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19: 716-723.

Avery T and Burkhart H. (1994). Forest measurements, fourth edition. *McGraw Hill*, Boston, Massachusetts, pp 408.

Blackard J. (2011). Personal communications. December 5, 2011.

Brown S and Barber J. (2012) The Region 1 Existing Vegetation Mapping Program (VMAP) Flathead National Forest Overview; Version 12. Rep. US Forest Service Region 1, 12-34. pp 5.

Crookston N and Finley A. (2008) yaImpute: An R package for KNN imputation, *Journal of Statistical Software*, 23(10): 1-16.

Fry J, Xian G, Jin S, Dewitz J, Homer C, Yang L, Barnes C, Herold N, and Wickham J. (2011). Completion of the 2006 National Land Cover Database for the Conterminous United States, *Photogrammetric Engineering and Remote Sensing*, 77(9): 858-864.

ESRI (2010 a). ArcGIS desktop help - How Principal Components work, accessed online: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Principal_Components_works/009z000000qm000000/, last accessed 5/6/2014.

ESRI (2010 b). ArcGIS desktop help - How ISO clustering works, accessed online: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Iso_Cluster_works/009z000000q8000000/, last accessed 5/6/2014.

FIA (2014). FIA DataMart 5.1, Accessed online: http://apps.fs.fed.us/fiadb-downloads/datamart.html , last accessed 5/6/2014.

Foody G, Mathur A, Sanchez-Hernadez C, and Boyd D. (2006). Training set size requirements for the classification of a specific class, *Remote Sensing of Environment*, 104: 1-14.

Haralick R, Shanmugam K, and Dinstein I. (1973). Texture features for image classification, *IEEE Trans.Syst.Man. Cybern.*, 3: 610-621.

Hogland J, Billor N, and Anderson N. (2013 a). Comparison of standard maximum likelihood classification and polytomous logistic regression used in remote sensing, *European Journal of Remote Sensing*, (46): 623-640.

Hogland J, Anderson N, and Jones G. (2013 b). Function modeling: improved raster analysis through delayed reading and function raster datasets, Proceedings of the Council on Forest Engineering (COFE) Annual Meeting, Missoula MT, USA. Accessed online: http://www.fs.fed.us/rm/pubs_other/rmrs_2013_hogland_j001.pdf, last accessed 5/6/2014.

Hogland J and Anderson N. (2014). Improved analyses using Function Datasets and statistical modeling. Proceedings of 2014 ESRI international users conference Annual Meeting, San Diego, CA.

Jenkins J, Chojnacky D, Heath L, and Birdsey R. (2003). National-scale biomass estimation for United States tree species, *Forest Science*, 49(1): 12-35.

Mauck J, Brown K, and Carswell W. (2009). The National Map—Orthoimagery: U.S. Geological Survey Fact Sheet 2009-3055, 4 p.

Oswalt SN, Smith WB, Miles PJ, and Pugh SA. (2014). Forest Resources of the United States, 2012. Gen. Tech. Rep. WO-xxx (Review Draft), Washington, DC: U.S. Department of Agriculture, Forest Service, Washington Office.

RMRS (2012). RMRS Raster Utility, Accessed online: http://www.fs.fed.us/rm/raster-utility/, last accessed 5/6/2014.

SAS/STAT(R) (2010 a). 9.22 Users Guide, Accessed online: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect016.htm, last accessed 5/14/2014.

SAS/STAT(R) (2010 b). 9.22 Users Guide, Accessed online: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect013.htm, last accessed 5/14/2014.

Scott T, Bechtold W, Reams G, Smith W, Westall J, Hansen W, and Moisen G. (2005). Sample-Based Estimators Used by the Forest Inventory and Analysis National Information Management System. In The enhanced forest inventory and analysis program — national sampling design and estimation procedures. Edited by Bechtold WA and Patterson PL. USDA For. Serv. Gen. Tech. Rep. SRS-80, pg. 43–67.

U.S. Department of Agriculture Forest Service. (2012). Forest inventory and analysis national core field guide: field data collection procedures for phase 2 plots. Version 6.0. Vol. 1. Internal report. Accessed online: http://www.fia.fs.fed.us/library/field-guides-methods-proc/docs/2013/Core%20FIA%20P2%20field%20guide_6-0_6_27_2013.pdf, last accessed 5/6/2014.

USGS (2012). Earth explorer, Accessed online: http://earthexplorer.usgs.gov/, last accessed 5/6/2014.

Wilson T, Lister A, and Riemann R. (2012). A nearest-neighbor imputation approach to mapping tree species over large areas using forest plots and moderate resolution raster data, *Forest Ecology and Management*, 27:182-198.

## *Tables*

**Table 1.** Potential explanatory variables and a short description of their assumed relevance to the probabilistic classification of visually identifiable patterns in NAIP imagery.

| Variable | Description |
|---|---|
| PC-1 | The score values of the first principal component, which represents an overall brightness metric of NAIP CIR imagery. |
| PC-2 | The score values of the second principal component, which represents an overall measure of homogeneity in surrounding spectral values (window size of 3 pixels by 3 pixels) of NAIP CIR imagery. |
| PC-3 | The score values of the third principal component, which highlights an inverse relationship between red and near infrared reflectance. This relationship is often attributed to vegetation and represents a vegetative index value. |
| GLCM-1* | Second order texture values of the brightness PCA component. These values are assumed to capture different aspects of brightness patterns within a defined analysis window (3 pixels by 3 pixels). |
| GLCM-3* | Second order texture values of the vegetation index PCA component. These values are assumed to capture different aspects of vegetation index patterns within a defined analysis window (3 pixels by 3 pixels). |

*GLCM transformations consist of 13 different calculations (Hogland and Anderson 2014) using horizontal $[0 \quad 1]$ and vertical $\begin{bmatrix}1\\0\end{bmatrix}$ offsets. GLCM transformations include: Contrast (CON), Homogeneity (HO), Dissimilarity (DIS), Entropy (ENT), Energy (EN), Angular Second Moment (ASM), Mean, Variance (VAR), Covariance (COV), Correlation (COR), Max Probability (MaxP), Min Probability (MinP), Range (Rng).

**Table 2.** Visually identifiable patterns and associated labels used to estimate forest characteristics.

| Pattern ID | Label | Category |
|---|---|---|
| 1 | Bare soil | BS |
| 2 | Crop | CRP |
| 3 | Dead trees | DEAD |
| 4 | Dead grass | DGRA |
| 5 | Live grass | GRA |
| 6 | Hardwood | HDW |
| 7 | Pavement | PV |
| 8 | Rock | Rock |
| 9 | Shrubs | SHR |
| 10 | Shadow | Shadow |
| 11 | Aspen | TAA |
| 12 | Pinyon and juniper | TPJ |
| 13 | Ponderosa pine | TPP-PP |
| 14 | Spruce/fir and Douglas fir | TSF |
| 15 | Water | WAT |

**Table 3.** Potential explanatory variables and a short description of their assumed relevance to estimating forest characteristics. For each Variable Transformation (column 1), PCA and probabilistic raster outputs were transformed to create potential explanatory variables.

| Variable Transformation | Description |
|---|---|
| SE | These explanatory variables (4) quantify the amount of shadow and tree, shadow and shrub, shadow and grass, and shadow and bare ground edge found within a focal extent of 78 x 70 pixels. The values for each variable are calculated by identifying locations that have a strong probability of being a shadow and a tree, shrub, grass, or bare ground visually identifiable class. These metrics are thought to be related to the number of trees found within a given FIA plot. |
| M | These explanatory variables represent mean values for a focal extent of 78 x 70 pixels (i.e., extent of FIA plot) calculated from raster PCA components and visually identifiable pattern outputs. In the case of the PCA Raster datasets, these explanatory variables (3) describe the actual spectral information before being transformed to visually identifiable patterns and are used to determine if visually identifiable patterns help to improve estimating forest community characteristics. For visually identifiable pattern outputs (15), these explanatory variables define the mean probability of a given pattern and are thought to be related to forest community species. |
| STD | These explanatory variables represent standard deviation values for a focal extent of 78 x 70 pixels (i.e., extent of FIA plot) calculated from raster PCA components and visually identifiable pattern outputs. In the case of the PCA Raster datasets, these explanatory variables (3) describe the variability in PCA scores within the extent of a FIA plot and are used to determine if variability in visually identifiable patterns help to improve estimating forest community characteristics. For visually identifiable pattern outputs (15), these explanatory variables define the standard deviation in probability of a given pattern and are thought to be related to forest community species and tree density. |
| M-STD | These explanatory variables represent mean values (focal extent 78 x 70) of the standard deviation values (focal extent 3 by 3) calculated from raster PCA components and visually identifiable pattern outputs. In the case of the PCA Raster datasets, these explanatory variables (3) describe the mean local neighboring variability in PCA scores within the extent of a FIA plot and are used to determine if variability in visually identifiable patterns help to improve estimating forest community characteristics. For visually identifiable pattern outputs (15), these explanatory variables define local neighboring variability in probability of a given pattern and are thought to be related to forest community species and average tree size. |
| GLCM* | These explanatory variables (252) represent second order texture values of PCA and probabilistic outputs. These values are assumed to capture different aspects of texture within a 78 x 70 pixels focal extent. |

*GLCM transformations consist of 7 of the 13 different GLCM calculations that can be performed within RMRS Raster Utility toolbar (Hogland and Anderson 2014) using horizontal and vertical offsets (see Table 1). These transformations include: Contrast (CON), Homogeneity (HO), Dissimilarity (DIS), Mean, Variance (VAR), Covariance (COV), and Correlation (COR).

**Table 4.** Tree species associated with forest communities used in the Uncompahgre National Forest.

| Community | Common Names | Latin Names |
|---|---|---|
| Aspen | Quaking Aspen | *Populus tremuloides.* |
| Fir | Douglas Fir, Alpine Fir, & Corkbark Fir | *Pseudotsuga menziesii, Abies lasiocarpa, & Abies lasiocarpa var. arizonica* |
| Juniper | Rocky Mountain Juniper & Utah Juniper | *Juniperus scopulorum & Juniperus osteosperma* |
| Pine | Lodge Pole Pine & Ponderosa | *Pinus contorta & Pinus ponderosa* |
| Pinyon | Pinyon Pine | *Pinus edulis* |
| Scrub Oak | Mountain Mahogany & Gambel Oak | *Cercocarpus spp. & Quercus gambelii* |
| Spruce | Engelmann Spruce & Blue Spruce | *Picea engelmannii & Picea pungens* |

**Table 5.** Eigen vectors obtained from a PCA using NAIP CIR images and rescaled first order standard deviation measure of texture for a 3 x 3 moving window. The first three components (PC1, PC2, and PC3) account for approximately 98.5% of the variation within the NAIP and texture parameters and can be interpreted as a brightness, texture, and vegetative index, respectively. Proportion of variation explained (%) for each component is identified in parenthesis following the component label.

| Parameter | PC1 (64.0%) | PC2 (23.9%) | PC3 (10.6%) | PC4 (0.8%) | PC5 (0.5%) | PC6 (0.2%) |
|---|---|---|---|---|---|---|
| NAIP Band1 (0.7-0.9µm) | 0.475 | 0.156 | 0.820 | -0.077 | 0.266 | 0.012 |
| NAIP Band2 (0.6-0.7µm) | 0.608 | 0.174 | -0.559 | -0.241 | 0.469 | -0.099 |
| NAIP Band3 (0.5-0.6µm) | 0.575 | 0.136 | -0.098 | 0.402 | -0.687 | 0.086 |
| First order texture Band1 | -0.170 | 0.446 | -0.051 | 0.762 | 0.415 | 0.133 |
| First order texture Band2 | -0.138 | 0.594 | -0.032 | -0.412 | -0.152 | 0.659 |
| First order texture Band3 | -0.163 | 0.613 | 0.040 | -0.155 | -0.203 | -0.728 |

**Table 6.** Candidate models used to predict visually identifiable patterns in the NAIP imagery. The top ranking model (Rank =1) has a ΔAIC value less than 2 and the fewest explanatory variables.

| ID | Rank | Explanatory Variables | DF | -2 Log Likelihood | AIC | ΔAIC |
|---|---|---|---|---|---|---|
| A | 1 | PC-1, PC-2 PC-3, H-HO-1, V-COR-1, H-COR-1, & H-HO-3 | 98 | 4551.675 | 4775.675 | 0.00 |
| B | 2 | PC-1, PC-2 PC-3, H-HO-1, V-COR-1, & H-COR-1 | 84 | 4597.256 | 4793.256 | 17.58 |
| C | 3 | PC-1, PC-2 PC-3, H-HO, V-COR-1, H-COR-1, & V-VAR-3 | 98 | 4572.48 | 4796.48 | 20.80 |
| D | 4 | PC-1, PC-2 PC-3, V-COR-1, & H-COR-1 | 70 | 4629.964 | 4805.244 | 29.57 |
| E | 5 | PC-1, PC-2 PC-3, H-HO-1, V-COR-1, H-COR-1, & V-CORR-3 | 98 | 4583.03 | 4807.03 | 31.35 |
| F | 6 | PC-1, PC-3, V-HO-1, V-COR-1, H-COR-1, H-CON-1 | 84 | 4615.458 | 4811.458 | 35.78 |
| G | 7 | PC-1, PC-3, V-HO-1, H-HO-1, V-COR-1, H-COR-1, V-VAR-1 | 98 | 4590.228 | 4814.228 | 38.55 |

**Table 7.** An example of candidate AGB models used to predict forest characteristics. In this example seven Aspen tree community AGB model AIC values were compared to select the best fitting AGB model. Top model selected (Rank =1) for each forest community had ΔAIC values less than 2 and the fewest explanatory variables.

| ID | Rank | Explanatory Variables | DF | SSE | AIC* | ΔAIC |
|---|---|---|---|---|---|---|
| A | 2 | PC_H_COR_1, PC_H_COV_2, PP_V_CON_12, PP_V_COR_5, & PP_V_HOM_13 | 77 | 12513.36 | 428.30 | 0.00 |
| B | 3 | SE_4, PC_H_COR_1, PC_H_COV_2, PC_MEAN_2, PP_H_COR_2,PP_V_CON_12, PP_V_COR_5, PP_V_HOM_13 | 74 | 11908.51 | 430.19 | 1.89 |
| C | 1 | SE_4, PC_H_COR_1, PP_V_CON_12, & PP_V_COR_5 | 78 | 13125.16 | 430.26 | 1.96 |
| D | 4 | PC_H_COR_1, PC_H_COV_2, PC_MEAN_2, PP_H_COR_2, PP_H_COV_7, PP_V_CON_12, PP_V_COR_5, & PP_STD_11 | 73 | 11886.53 | 432.04 | 3.74 |
| E | 5 | PC_H_COR_1, PP_V_CON_12, & PP_V_COR_5 | 79 | 14141.82 | 434.46 | 6.16 |
| F | 6 | PC_H_COR_1, PP_V_CON_12, PP_V_COR_5, & PP_STD_11 | 78 | 14021.81 | 435.75 | 7.45 |
| G | 7 | PC_H_COR_1, PC_H_COV_2, PP_V_CON_12, & PP_V_COR_5 | 78 | 14124.38 | 436.35 | 8.05 |

$*AIC = n * ln\left(\frac{SSE}{n}\right) + 2p$

**Table 8.** Fit statistics and variables of top AGB models select for each forest community.

| Community | Explanatory Variables | $R^2$ | RMSE |
|---|---|---|---|
| Aspen | SE-4, PC-H-CORR-1, PP-V-CONT-12, & PP-V-CORR-5 | 0.72 | 12.97 |
| Fir | PC-H-CORR-3, PP-V-MEAN-11, & PP-V-CORR-5 | 0.46 | 11.32 |
| Juniper | PP-H-CONT-5, PP-H-COV-13, PP-H-CORR-1, PP-H-CORR-12, PP-V-CONT-8, PP-V-COV-9, & PP-V-CORR-8 | 0.72 | 3.86 |
| Pine | PC-H-CONT-3 & PP-H-COV-14 | 0.13 | 12.88 |
| Pinyon | SE-3, PP-H-CONT-13, PP-H-DIS-2, PP-H-CORR-3, PP-H-CORR-12, PP-V-CONT-9, PP-V-MEAN-5, & PP-V-VAR-14 | 0.62 | 3.12 |
| Scrub Oak | PP-H-CONT-6, PP-H-VAR-10, PP-V-MEAN-10, & PP-V-CORR-5 | 0.30 | 3.99 |
| Spruce | SE-1, SE-3, PC-H-CONT-3, PC-H-DIS-1, & PC-H-CORR-3 | 0.66 | 12.78 |

***Figures***

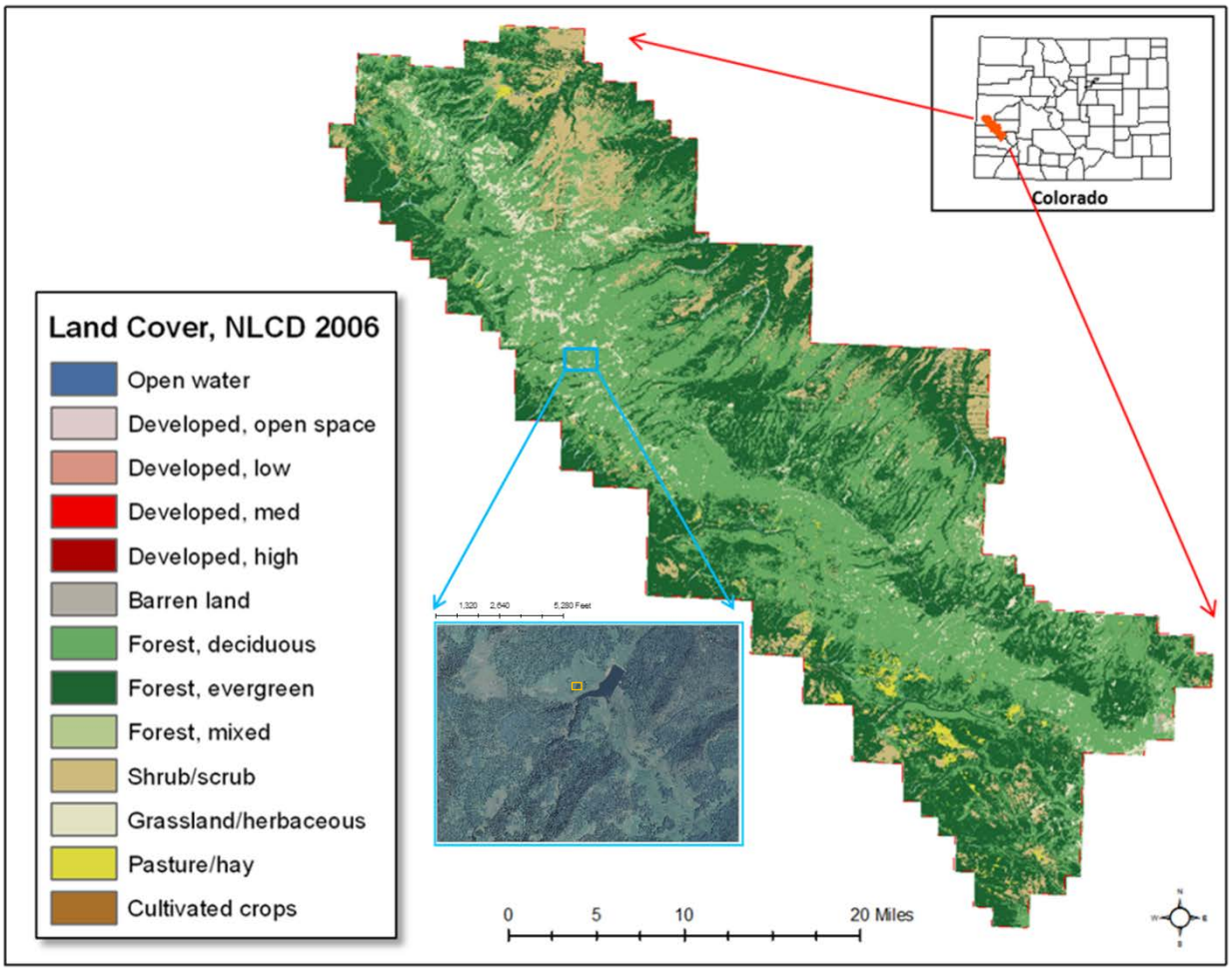


**Figure 1.** Location map of the Uncompahgre National Forest (UNF) and associated ecological/land cover communities taken from the National Land Cover Database (Fry et al. 2006). The boxed area in light blue identifies a subset of the UNF, displayed at a larger spatial scale, which is used to illustrate the spatial resolution of the data and associated explanatory variables of stage 1 and 2 classification and estimation models. The orange box within the blue subset portrays the spatial extent of a Forest and Inventory Analysis Program plot (not an actual plot location).

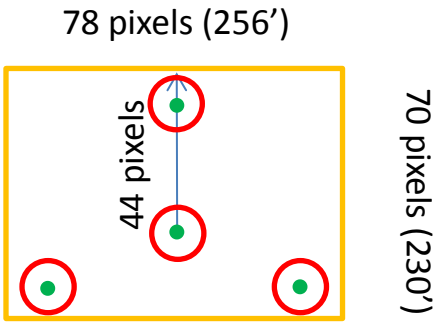


**Figure 2.** Spatial layout of an FIA plot and moving window used to summarize focal and GLCM values. Subplots (green dots) are 120 feet from the central plot at a bearing of 0, 120, and 240 degrees. The circumference of each subplot (red circles) has a radius of 24 feet and amounts to a land area of approximately 1/24 of an acre. The orange bounding box identifies the size of the moving window used to summarize predictive values derived from NAIP imagery.

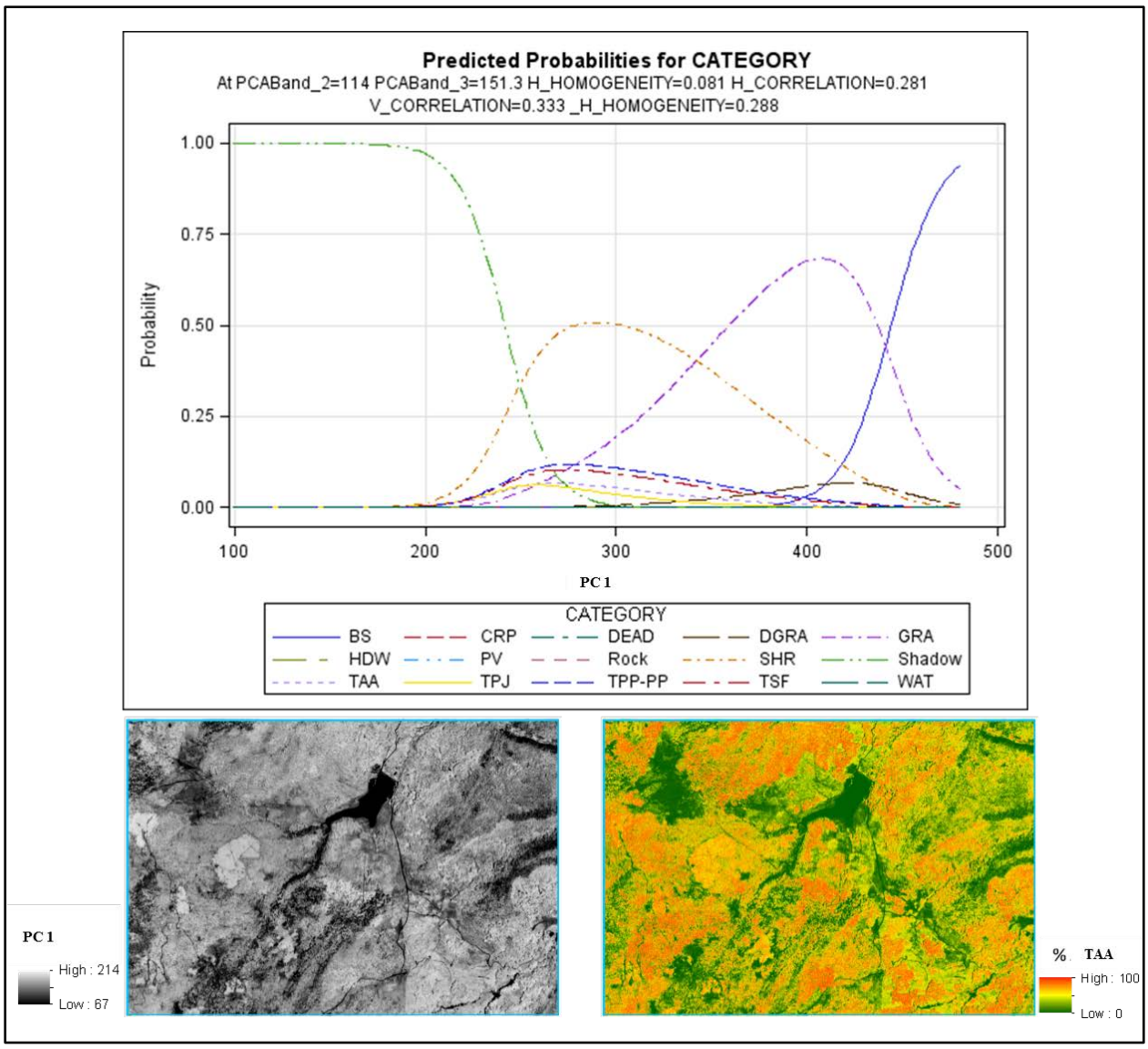


**Figure 3.** Probability distribution of top fitting PLR model in feature space (top graphic) and geometric space (bottom graphic). Note the change in probability of the visually identifiable pattern labeled TAA as the first component of the principal component analysis (PC 1) varies from small to large values (while holding all other variable at their mean values). While the probability of TAA changes gradually in feature space for varying values of PC1, in geometric space PC1 and corresponding TAA probability can change quite dramatically for neighboring pixels.

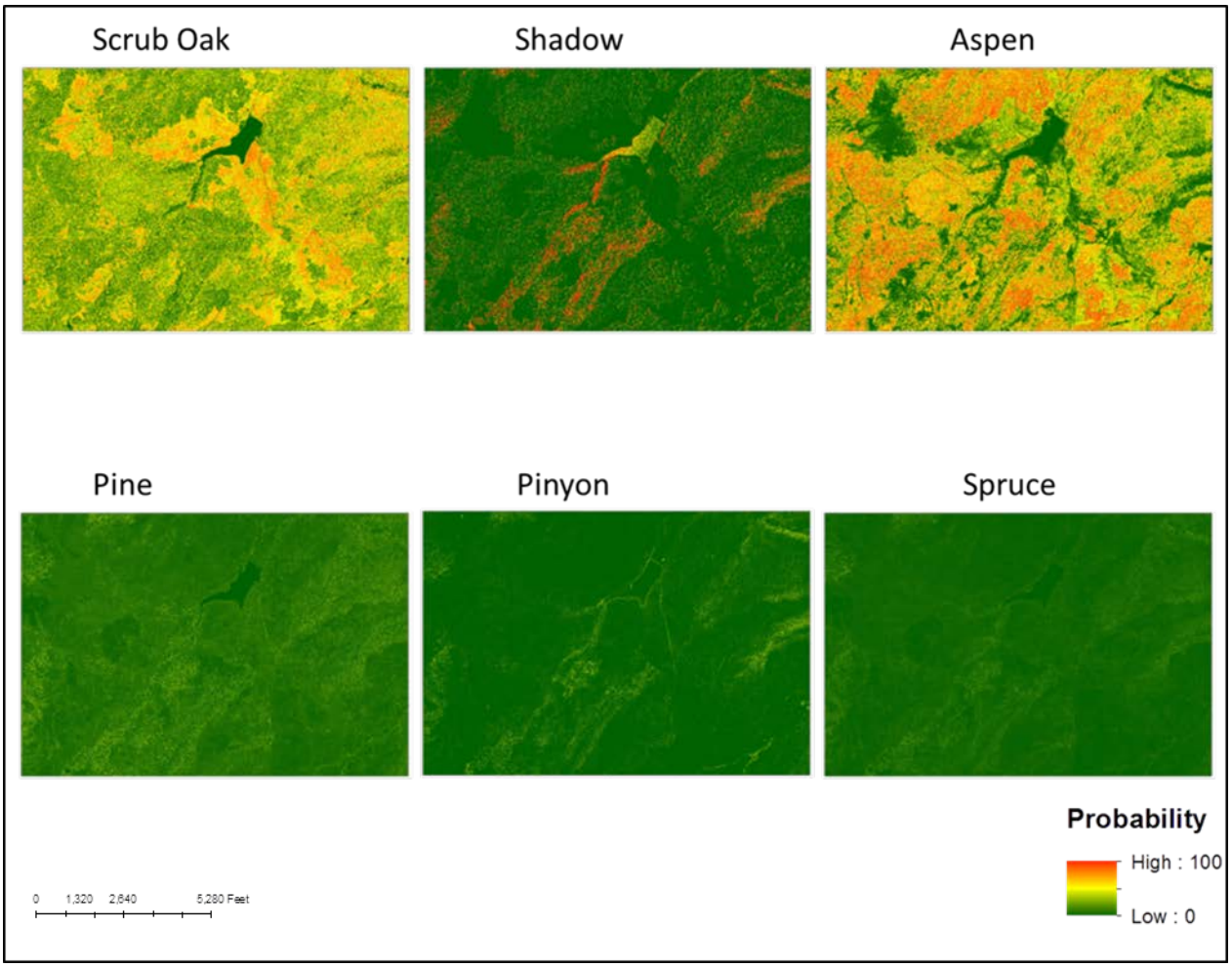


**Figure 4.** Example of six probabilistic surfaces that identify visual patterns thought to be related to forest characteristics in the NAIP imagery for a subset of the UNF (Figure 1). As pixel values transition from green to red, the probability associated with each NAIP pattern increases from 0 to 100%.

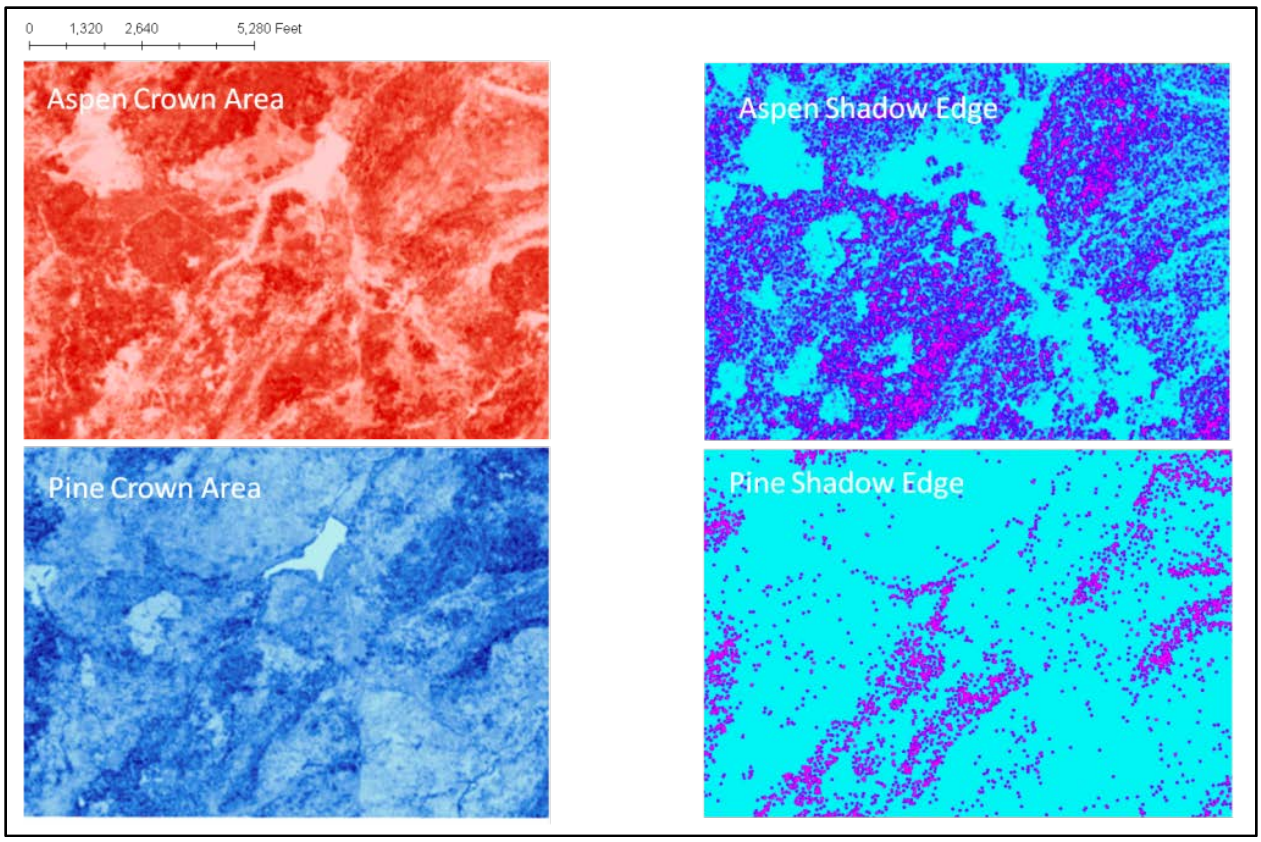


**Figure 5.** An example of four potential explanatory variables used to estimate community characteristics for the subset of the UNF identified in Figure 1.

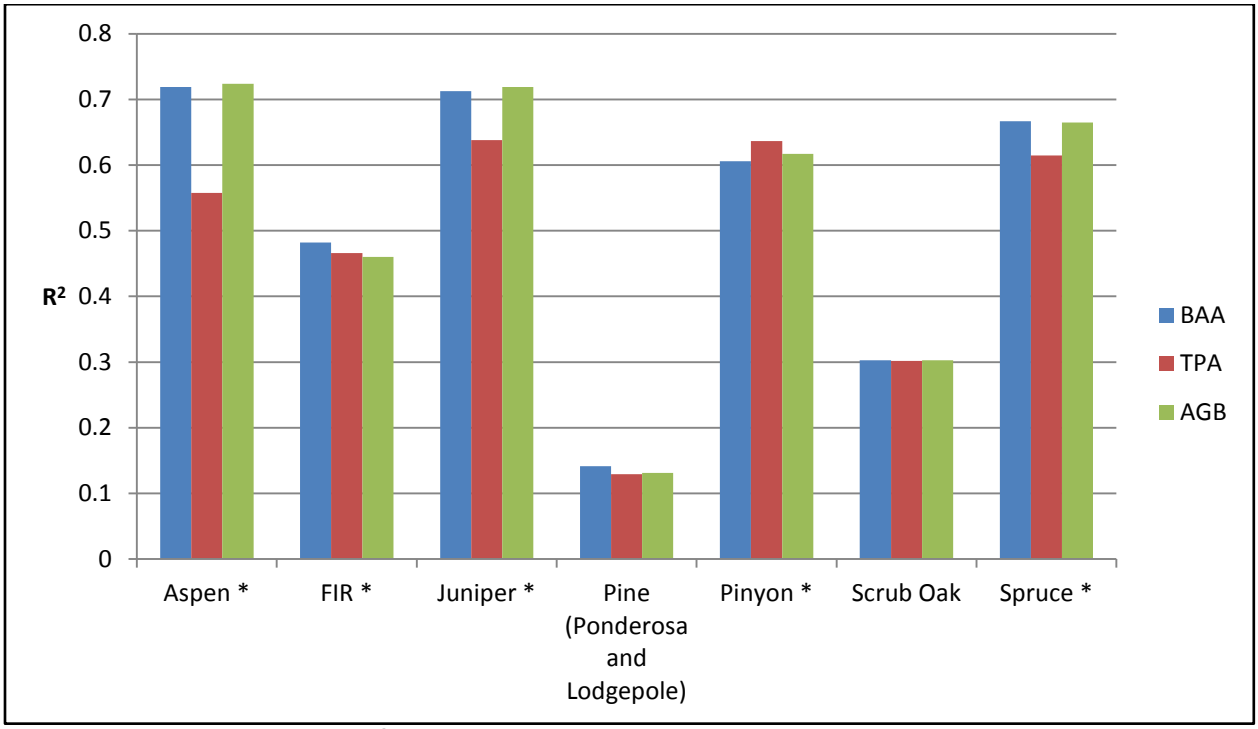


**Figure 6.** Model fit statistics ($R^2$) for forest community characteristics. The asterisks (*) next to forest community names denote communities most common in UNF.

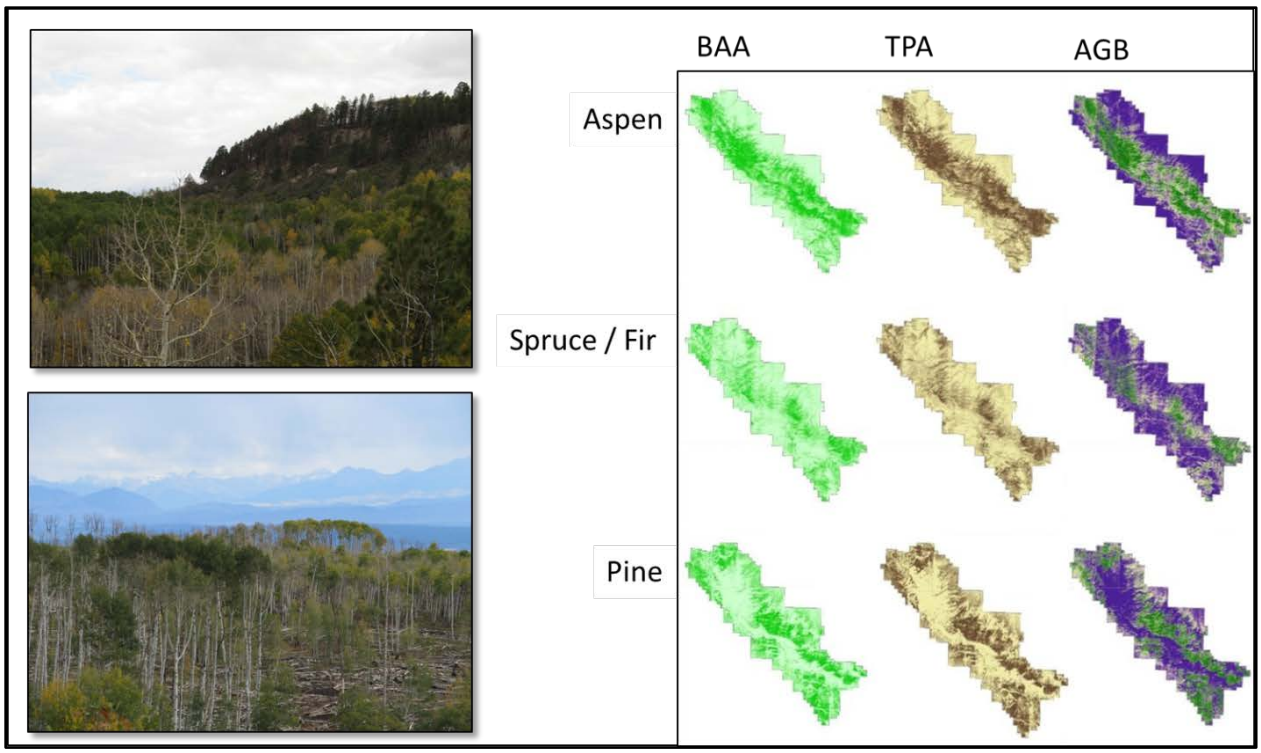


**Figure 7.** A subset of the outputs created from the second stage of our modeling process. Pictures on the left illustrate the heterogeneous nature of Aspen, Spruce/Fir, and Pine forest communities across a small subset of UNF. Raster surfaces on the right depict BAA, TPA, and AGB for these communities across the entire UNF.

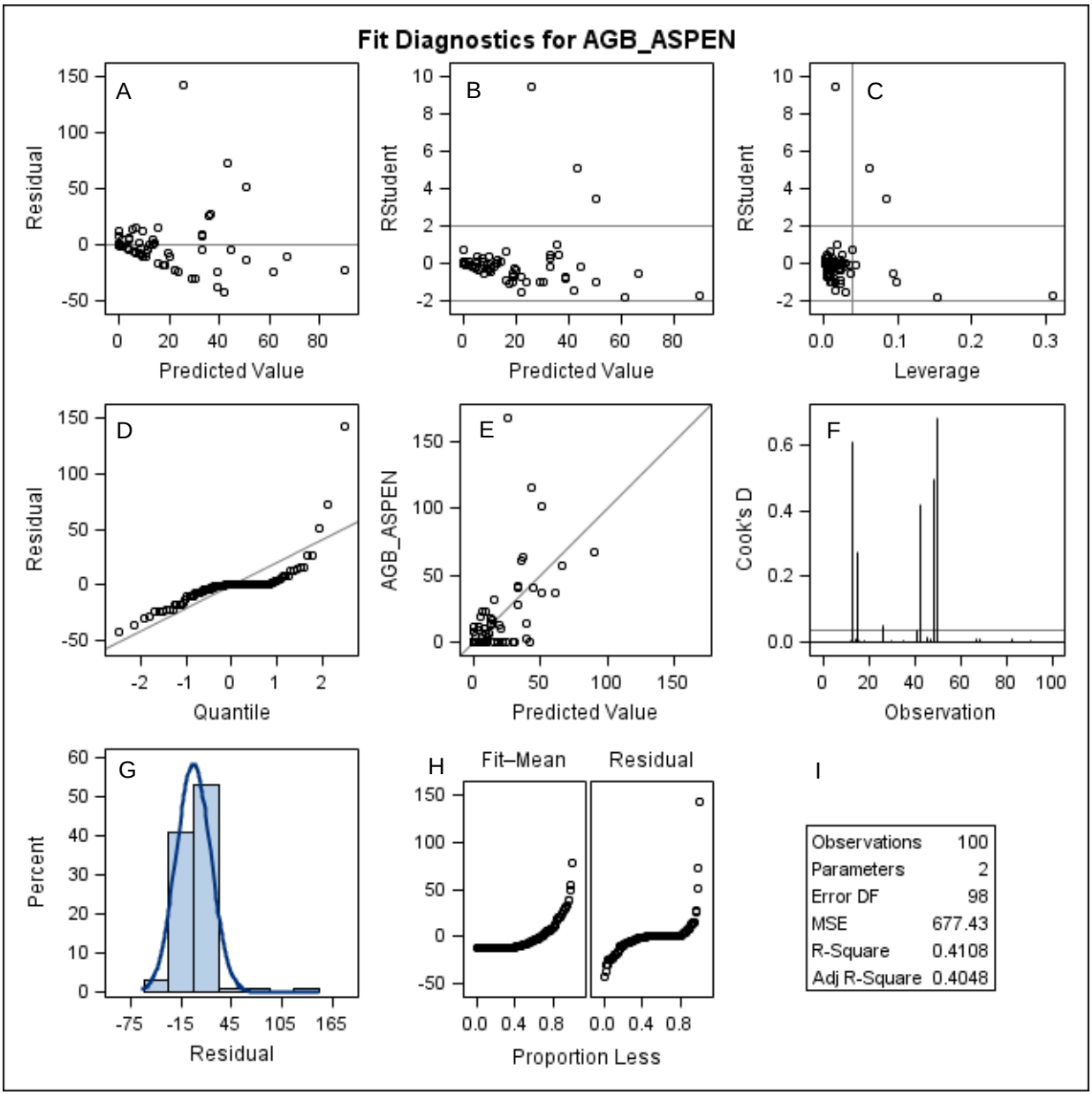


**Figure 8.** An example of Aspen AGB residuals calculated from an ordinary least squares regression using observed (response) and predicted (explanatory) values from the independent validation dataset. Regression slope and intercept estimates for this example were not statistically different than one and zero, respectively (p-value <0.001). Note the linear grouping of residual values in graphics A, B, and E. These observations had no measured AGB and illustrate the tendency of our models to predict small amount of AGB when no AGB is present.

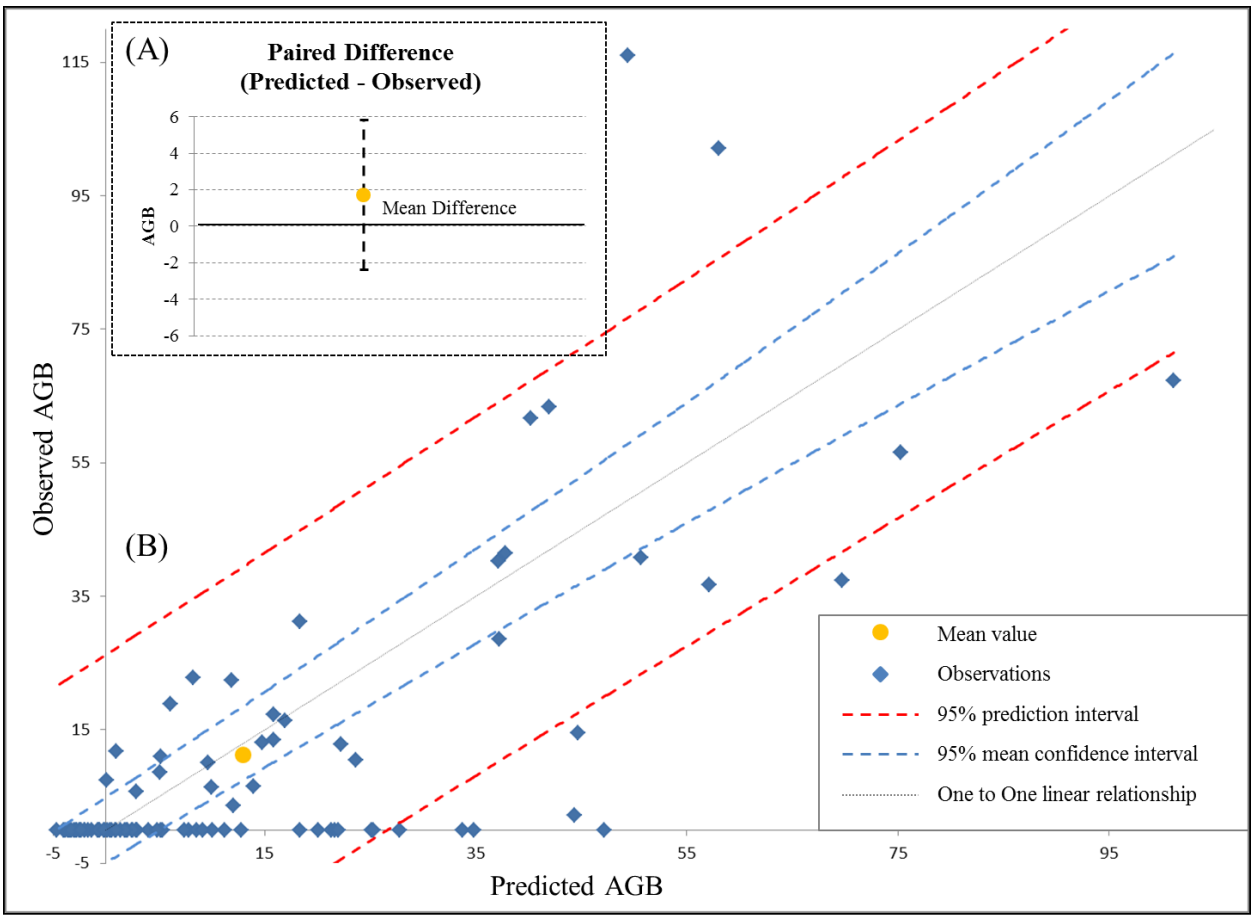


**Figure 9.** An example of observed versus predicted AGB values for Aspen forest community type derived from our validation dataset. While the vast majority of observations from the independent validation dataset fall within the bounds of the prediction interval, individual predicted values can vary substantially from observed values. However, when aggregated together (means shown by orange circles) the predicted mean estimate is almost identical to the observed mean estimate of AGB (close to zero in inset A and close to the one-to-one line in B) and statistically falls within the bounds of the 95% confidence interval for the mean (black confidence intervals in inset A and blue dashed lines in B).