Predicting tree species composition at pixel-scale

Raymond L. Czaplewski[1]
Michael Hoppus[2]
Andrew Lister[3]
January 26, 2005

Detailed data on tree species drive models that predict risk of insect and disease mortality in forest stands and simulation models for future stand conditions. Application of such models in geospatial analyses requires these data for millions of remotely sensed pixels. However, the vast majority of remotely sensed thematic maps predict a few categories of stand conditions, such as forest type and stage of stand development. Even if the inherent inaccuracies in remotely sensed predictions are ignored, there is considerable variability in tree composition within each category. There is growing interest in *k*-Nearest Neighbor (*k*-NN) imputation as an alternative to supervised classification of remotely sensed data. *k*-NN starts with a set of training sites j, $10^2<j<10^3$, within the target geographic area. One attractive set of training sites is the field plots measured by the USDA Forest Service's Forest Inventory and Analysis program. For each pixel i, $10^5<i<10^7$, which are outside of the training set, *k*-NN finds $1 \leq k$ training sites that are "close" to the i^{th} pixel within a feature space formed from remotely sensed and other geospatial data. Then detailed field measurements from those k training sites are used to impute, or predict, the same type of detailed field data for that i^{th} pixel. This imputation is separately repeated for each and every pixel in the full target area. The outcome is detailed predictions of tree species, tree size composition and other field measurements for each pixel. The accuracy of *k*-NN predictions strongly depends upon the distance metric used to measure "closeness" in this feature space, and there are numerous alternatives for that measure. This paper presents and evaluates a new measure that transforms a high-dimensional remotely-sensed feature space into a new space that is optimized to fit a high-dimensional response space, namely tree-level composition at the pixel scale. The advantages of this approach include highly efficient prediction algorithms for risks to forest health and forecasts of future conditions at the pixel scale.

Efficient forest management requires accurate maps of risks to forest health. Some risk models use detailed data on the distributions of tree species and tree sizes in a stand, which are best directly measured in the field. In a more perfect world, these data would be available for every stand. Therein lays the problem. While the field measurements required for some risk models are feasible at a small number of sample sites, they are seldom feasible in all stands across a large geographic area. Therefore, the challenge is to extrapolate field measurements from a sample to full wall-to-wall map coverage.

[1] USDA Forest Service, Rocky Mountain Research Station; 2150 Centre Ave. Bldg. A; Fort Collins, CO 80526; 970-295-5973; rczaplewski@fs.fed.us
[2] USDA Forest Service, Northeastern Research Station, 11 Campus Blvd. Suite 200; Newtown Square, PA 19073-3294; 610-557-4039; mhoppus@fs.fed.us
[3] USDA Forest Service, Northeastern Research Station, 11 Campus Blvd. Suite 200; Newtown Square, PA 19073-3294; 610-557-4038; alister@fs.fed.us

This is not a new challenge. Forest stand mapping with aerial photographs has been common for 50 years. An photo-interpreter classifies each stand into a single category for forest type (e.g., aspen-birch association), stocking (e.g., 70-100% crown closure), and stage of stand development (e.g., poletimber). In reality, each delineated stand has internal variability (e.g., interior, edge, inclusions, gaps, gradients), and a single category containing numerous stands has even greater variability. However, it is not reasonable to delineate and classify stands into smaller pieces or more detailed categories. This variability within each category makes accurate risk modeling problematical.

Digital classification with remotely sensed data has greatly improved the efficiency of mapping stand conditions. However, today's technology still follows the 1950's paradigm: classification into a single category of broad forest conditions with considerable variability within each category. This paradigm remains frustrating to entomologists and pathologists who predict and map expected forest health risks based on more detailed stand conditions.

Another shortcoming of modern remote sensing methods is in the type of training data used in digital classification algorithms. Remote sensing technologists favor homogeneous training areas, such as the interior of an undisturbed even-age stand. They avoid heterogeneous stands or stand edges, even though these messy conditions can dominate many landscapes across the eastern United States.

An alternative to traditional classification of remotely sensed imagery is *k*-Nearest Neighbors (*k*NN) imputation. Simply stated, *k*NN predicts (imputes) the unknown conditions of an unmeasured pixel by averaging the known measurements of *k* pixels, where *k* often ranges from 1 to 20. kNN need not classify pixels into categories. Rather, kNN predicts a tree-list for every pixel in the image using a sample of training data that has a tree-list from a field cruise.

The accuracy of *k*NN depends on which measured pixels are used to predict field data for each unmeasured pixel. That is where the "Nearest Neighbors" part comes in. The assumption is that pixels with similar remotely sensed data will have similar field measurements. *k*NN methods measure similarity between two pixels as the Euclidean distance in a multidimensional "feature space," which is defined by multivariate remotely sensed data. For example, we used Spring and Fall Landsat imagery, each with six spectral bands; and elevation, slope and aspect. Every pixel *i* has a precise location in this15-dimensional feature space. To predict the field measurements for pixel *i* with unknown stand conditions, the *k*NN algorithm searches among the training sample of *n* pixels, for which field measurements are known. The algorithm finds the *k* pixels among the sample of *n* pixels that have the shortest distance to pixel *i* in this15-dimensional feature space of remotely sensed data. The algorithm then averages the field measurements from these *k* "nearest neighbors" and assigns those averages to pixel *i*. The algorithm is repeated for each pixel in the entire mapped area.

The accuracy of the *k*NN method can be increased by transformations of the remotely sensed feature space so that pixels with similar stand conditions are closer together in the transformed space. This has often been accomplished by simple *ad hoc* methods. For

example, each of the 15 layers of remotely sensed data could be normalized to have a mean of zero and a variance of one. The Mahalanobis transformation goes one step further by making each of the 15 layers statistically uncorrelated with each other. Each of the 15 remotely sensed layers can be scaled by a constant to reduce prediction residuals in a test dataset, either by systematic experimentation with different sets of fixed weights, or with a nonlinear optimization algorithm. Canonical correlation transforms the multivariate remotely sensed feature space to maximize the correlation with multivariate field measurements from a training sample.

This paper explores a very simple alternative to the usual transformation strategies. The training sample is used to fit linear or nonlinear multivariate regression models that predict stand-level total basal area for each of 19 major tree species using multivariate remotely sensed measurements as predictor variables. Then the models are applied to all pixels in the study area. This produces a new 19-dimension feature space, each with the units of predicted basal area per hectare for one of the 19 major tree species. Finally, the *k*NN algorithm is applied in this new feature space. The assumption is the optimization function in regression will produce an "optimal" feature space in which the *k*NN algorithm can operate. To the best of our knowledge, this very simple idea has not previously appeared in the forestry literature.

We used 717 field plots that cover the entire State of New Hampshire. We simultaneously compared 86 different variations of typical transformations plus the new transformations based on regression models. We discuss results for the 24 "best" models among these 86. This is a single case study, so we do not make generalizations about performance of any of these methods in other study areas.

We randomly selected 500 plots as training data, and the remaining 217 plots were used to access accuracy. Both of these partitions provide sample estimates for the entire population.

The models considered are based on permutations of the following options:

(1) The *k* in *k*NN is 1, 5, or 15 nearest neighbors.

(2) A single donor plot is randomly selected within a cell of 5 or 15 of the nearest neighbor plots. This is based on a imputation method by Brick et al. (2004), but it has not appeared in the forestry literature on *k*NN methods.

(3) Basal area measurements or predictions are modified with the Hellinger transformation, which converts abundance by tree species to the fractional species composition of a site. This approach is recommended Legendre and Legendre (2003) for analyses of plant communities

(4) In addition to the 19-dimension basal area regression feature space, a 20th dimension is added for predicted total pixel-level basal area. Since the unit of measure for all 20 dimensions is basal area per hectare, this option puts approximately 50% of the importance on total stand basal area, and the remainder on basal area per hectare for each of the 19 major tree species.

(5) Feature space defined by the Mahalanobis transformation.

(6) Each dimension of the feature space is transformed into rank order statistics. The number of ranks equals the number of plots in the training data set (500 in our case). This is a self-scaling, nonparametric approach in that the original range of data values do not effect the ranks.

(7) Canonical correlation transformation.

(8) Multivariate multiple regression transformation

(9) Nonlinear binary tree regression, separately applied to each of the 19 manor tree species.

(10) A binary-tree nonlinear classifier was used with the 15 remotely sensed layers to group pixels into major forest types. For each pixel without field measurements, *k* training plots where randomly selected within the same remotely sensed category. If *k* is very large, this is similar to assigning a stratum mean to each pixel in that remotely sensed stratum.

(11) Each dimension of a transformed feature space is weighted by an index of its predictive strength. In canonical correlation, it is the eigenvalue for that canonical variate. In a feature space predicted from regression models, it is the correlation between measured and predicted BA by tree species from the training sample

(12) A 19-dimension *k*NN feature space is formed based on field measurements of basal area by the 19 major tree species in the sample of 217 test plots. This feature space can not be applied to the entire map, but it is useful to compare some of the above options in a “best-case” setting.

Accuracy is evaluated with the random sub-sample of 217 field plots. Accuracies are evaluated at both the pixel- and population-levels. However, this evaluation is complex because stand basal areas are predicted for 19 different species, and assessments at both the pixel- and population-levels are relevant in forestry applications.

Pixel level prediction accuracy is the squared difference between the predicted and true BA summed over all 19 major tree species. The residuals represent both the variability of predictions and their overall squared bias. There are 19 different residuals for each pixel in the test dataset, one for each major tree species. A 19x19 covariance matrix is computed for residuals in the test data. We measure the overall magnitude of these multivariate prediction residuals with the Wilk's generalized variance, which is a scalar value that quantifies the "volume" of a covariance matrix based on its matrix determinate.

Population level accuracy is measured by the Euclidean distance between each of 217 quantiles from the true and predicted BA's for each of the 19 tree species. This gauges the distribution of basal area per hectare by tree species within the entire population.

In the "best case" setting in option (12), pixel-level prediction errors are lowest for *k*=5, and errors are somewhat larger for *k*=1 and *k*=15. However, the agreement in basal area distribution by tree species at that population level is greatest with *k*=1; disagreement is about 50% worse with *k*=5, and about 100% worse with *k*=15. Therefore, selection of the *k* value in *k*NN can force a compromise between accuracy in predicting conditions at the pixel level v. the population level.

The most accurate options with remotely sensed data at the pixel-level included the rank transformation (12) with *k*=5 and *k*=15; multivariate multiple regression (8) and the Mahalanobis (5) transformations with *k*=5 and *k*=15; multivariate multiple regression (8) with *k*=1 and a neighborhood cell (2) having 5 nearest neighbors; the nonlinear binary tree regression models (9) and multivariate multiple regression (8) without *k*NN imputation; and canonical correlation (7) with *k*=15. All these options had very similar accuracies at the pixel level. Other options were noticeably less accurate.

These same options performed relatively well for species-specific basal area distributions at the population level. However, multivariate multiple regression (8) with *k*=1 in a neighborhood cell having 5 nearest neighbors (2) was notably more accurate. Given these particular accuracy metrics, this latter option performed best in this very limited case study.

Classification of pixels into major forest type categories (12) was not as accurate as these other options. This suggests that *k*NN can produce better results than assigning the stratum means of field measurements to all pixels in that remotely sensed stratum. Also, the Hellinger (3) transformation and weighting dimensions of the feature space by goodness-of-fit statistics (11) did not improve accuracies noticeably.

*k*NN methods can offer a several advantages compared to regression models alone. *k*NN better preserves the pixel-level covariance structure and population-level distributions among tree species, tree sizes and stocking, especially when *k*=1. Also, a risk model can be run once for each of the *n* field plots in the training sample. Since *k*NN associates each pixel with *k* field plots among those *n* plots, the model predictions from those plots form the prediction for the pixel. It is not necessary to run the risk model separately for each pixel in the entire mapped area, as would be the case with regression models alone.

In conclusion, regression transformations for *k*NN imputation in this one case study produced incremental improvements over more familiar *k*NN methods, such as the Mahalanobis and canonical correlation transformations. The improvement was most notable for population-level distributions rather than pixel-level accuracy.