MERGING RANDOM FOREST CLASSIFICATION WITH AN OBJECT-ORIENTED APPROACH FOR ANALYSIS OF AGRICULTURAL LANDS

J. D. Watts[a] and R. L. Lawrence[b].

[a]Spatial Sciences Center,Montana State University, Bozeman, USA, - jennifer_watts@hotmail.com
[b]Spatial Sciences Center, Montana State University, Bozeman, USA, - rickl@montana.edu

Working Group VII/4

KEY WORDS: land use, agriculture, Landsat imagery, segmentation, Random Forest

ABSTRACT:

Machine learning algorithms recently have made major advances, with decision tree classifiers gaining wide acceptance. Boosting and bagging of decision trees have added to the predictive capabilities of these approaches. Object-oriented (O-O) analyses have been developed during this same period, offering important improvements in classification over pixel-based approaches under certain conditions. Classification algorithms for O-O approaches, however, have been fairly limited and generally have not incorporated new statistical approaches used for pixel-based classifications. One of the most promising new classification algorithms is Random Forest (Breiman-Cutler) classification (RF). We incorporated RF into an O-O classification of Landsat-based imagery for mapping agricultural lands in north-central Montana, USA. The Definiens multi-resolution segmentation algorithm was used to generate field-based objects. RF was used to classify land management (tillage, conservation reserve, crop/fallow) based on reference data from >400 field sites. Object-based attributes included factors such as average spectral response, spectral variability, texture, and shape characteristics. Accuracy was assessed using "out-of-bag" estimates in RF. This classification approach was able to efficiently and accurately merge RF with an object-oriented approach for improved classifications.

1. INTRODUCTION

Advanced image classification algorithms are becoming increasingly popular within the remote sensing community. These include, but are not limited to, boosting and/or bagging-based classification and regression trees (CART) (Lawrence et al., 2004; Lawrence and Wright, 2001; Baker et al., 2001) and the CART-based Random Forest (RF) algorithm (Lawrence et al., 2006; Ham et al., 2005). These classification techniques have been utilized primarily on a per-pixel basis, in spite of advancements in object-oriented (O-O) image segmentation and analysis.

Object-based image analysis moves beyond the somewhat disconnected process of analyzing individual data points within a landscape by grouping together pixel-regions according to spectral and spatial similarity (Navulur, 2007). The resulting image objects serve as integrated entities that exhibit an intrinsic scale and are composed of structurally connected parts or pixels (Hay et al., 2003).

The RF classification algorithm is superior to many tree-based algorithms, because it lacks sensitivity to noise and is not subject to overfitting. Some studies have suggested that RF is "unexcelled in accuracy among current algorithms" (Breiman and Cutler, 2005). RF has also outperformed CART and similar boosting and bagging-based algorithms (Gislason et al., 2006; Pal, 2003). This algorithm uses bagging to form an ensemble of classification trees (CART-like classifiers) (Breiman, 2001; Gislason et al., 2006). Bagging, or bootstrap aggregating, forms multiple training sets by sampling from a primary data set with replacement (Breiman and Cutler, 2005). Bagging is

advantageous as it improves model stability; the model's predictive ability increases as data over-fitting is avoided. RF is distinguished from other bagging approaches in that at each splitting node in the underlying classification trees, a random subset of the predictor variables is used as potential variables to define the split.

RF utilizes the Gini index of node impurity (Breiman et al., 1998) to determine splits in the predictor variables that result in the greatest classification accuracy. Tree "branches" are split in a manner that reduces the uncertainty present in the data and hence the probability of misclassification. Ideal branch partitioning, or a Gini value of zero, occurs when only one class is represented at each terminal node. The bagging and splitting process continues until a "forest", consisting of multiple trees, is created. Classification occurs when each tree in the forest casts a unit vote for the most popular class (Breiman, 2001). This results in a classified output determined by a plurality vote. Unlike CART analysis, trees in RF are not pruned. Pruning is not needed as each classification is produced by a final forest that consists of independently generated trees created through a random subset of the data, avoiding over fitting (Breiman, 2001).

Another advantage to RF is given by an internal accuracy measure that makes it unnecessary to partition reference data into separate sets for training and validation. All available reference data, therefore, can be used to develop the predictive model. Test set accuracy is estimated in RF by running out-of-bag (OOB) samples (a subset of the training data that was not included in the bootstrap for a particular tree) down through a tree as a form of cross-validation. A study comparing OOB accuracy assessments with the traditional error matrix approach for resulting rangeland classifications reported that the OOB estimates were within 3% of the independent accuracy assessments, with most less than 1% apart (Lawrence et al., 2006). The authors cautioned, however, that OOB estimation is only reliable given an absence of bias in the reference data. An

RF package is currently available in both R and S-Plus (Insightful) statistical packages.

Some studies have utilized advanced classification algorithms within O-O analysis. Classification algorithms within the popular Definiens O-O platform are currently limited to a nearest neighbor (NN) classification that utilizes fuzzy logic and a membership function-based classification (Navulur, 2007). A decision tree classification was used in an object-based analysis of IKONIS imagery for forest inventories, resulting in producer's accuracies ranging from 81-100% and 86% total accuracy (Chubey et al., 2006). The study suggested that advanced classifiers with the ability to process an extensive number of inputs are required to take full advantage of the rich set of data that can be derived through O-O analysis. Another study used a CART-based classification to map Benthic coastline habitats with 74% overall accuracy (Green and Lopez, 2007). Other approaches utilizing advanced classification techniques within O-O analyses have included the application of CART-based rules to increase K-NN classifications (Yu et al., 2006) and the incorporation of genetic algorithm feature selection within neural network classifications (Van Coillie et al., 2007). Classification and regression tree algorithms are advantageous as they utilize advanced statistical techniques to produce in many cases a more accurate classification model; model rule sets can then be readily incorporated into the Definiens Developer (formerly Professional) O-O program to produce an object-based classification map. O-O studies utilizing the RF classifier have not been identified by the authors at this time. This might be due to a general lack of knowledge pertaining to the RF model within the O-O community and the inability to generate apparent rule sets that can be taken into the Definiens software for mapping the classification results.

We applied RF to field-based image objects derived from moderate resolution Landsat TM and ETM+ imagery in an attempt to identify accurately agricultural management practices, namely no-till (NT) and conservation reserve (CR). Field vegetative status was also determined, as this information might be used to determine multi-year crop and fallow patterns for cropping intensity purposes.

2. METHODS

Our focus was on mapping dry land cropping practices within north central Montana. This semi-arid region is known for its production of dry-land wheat. Area farmers have been encouraged to implement conservation practices, such as NT and CR, to increase soil organic carbon (Fawcett and Towery, 2002). The implementation of continuous cropping, or exclusion of summer fallowing, also has been suggested. Summer fallow is when cropland is left un-vegetated for a growing season to increase soil moisture storage.

Field management data were collected early June 2007 for locations randomly selected throughout the region. The resulting cropland data set included information for 78 NT-fallow, 138 NT-cropped, 48 tilled-fallow, 148 tilled-cropped, and 113 CR field sites. The actual number of field sites utilized within the model-building process was scene-dependent due to cloud masking and missing pixel information resulting from ETM+ scan-line gaps.

Landsat image pairs (path 39; rows 26, 27) were obtained for 15 May (Landsat 5 TM) and 11 August (Landsat 7 ETM+) 2007. Geometric correction techniques were used to ensure that the images were properly aligned within geographic space, followed by cloud and shadow masking to remove contaminated pixels. Image data were then converted to exoatmospheric reflectance to minimize between-image differences due to earth-sun distance and solar angle (Chander et al., 2007; SDH-L7, 2006). Normalized Difference Vegetation Indices (NDVI), representative of relative photosynthetically active vegetation densities (Tucker and Sellers, 1986), and the Tasseled Cap components associated with soil brightness, vegetation greenness, and surface wetness (Crist et al., 1986; Huang et al., 2002) were also included as predictors. It was thought that the addition of these indices might better allow for node splitting within the model. A non-cropland mask was applied to remove water bodies, urban and public lands, transportation networks, and rangeland.

Vector-based image objects representing parcel management strips and within-strip sections of spectral and textural similarity were generated through the multi-resolution O-O segmentation algorithm (Benz et al., 2004). The within-strip segmentation was used to reduce the inclusion of both crop and bare soil within an image-object. A strip-based segmentation was determined to be suitable for tillage and CR classifications, as it was unlikely that these management types would vary within field-based boundaries. Vector information representing taxable field parcels was also included within the segmentation process, to ensure that generated objects were constrained within ownership boundaries (Figure 1).

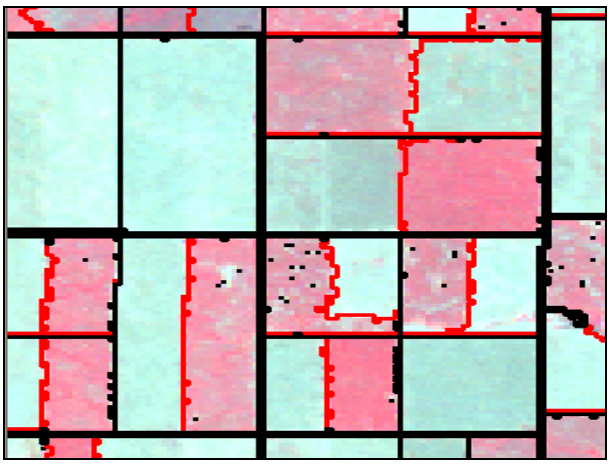

Figure 1. Object segmentation results for the field strip level (red vector lines). Black parameter lines represent taxable field boundaries.

Resulting object-based attribute data were imported into the randomForest package (S-PLUS®) to generate classification models for NT and till, CR and cropland, cropped and fallow. These included spectral, textural, and neighborhood object-based parameters. Initial forest models were built using 500 generated classification trees, the default number. Model tree adjustments were based on an analysis of model error as influenced by the number of RF trees. Model classification matrices and associated class accuracies were determined through the internal OOB accuracy assessment (Breiman, 2001). Data from either image dates, or predictor parameter sets utilizing data from both image dates, were examined in the generation of class models. A May TM pixel-based tillage model also was examined, in addition to the object-based models, to ascertain the effect that object-based textural and

neighborhood parameters might have in improving tillage accuracy.

Class predictions were exported and joined with the existing vector objects (in .shp format) according to field identification numbers, within a GIS platform (Figure 2). This allowed for an efficient way to examine spatial relationships between cropland management class predictions, spectral image data, and various other data sets.

Figure 2. The vector-based cropped and fallow classification layer, overlaying a Sept. 2007 ETM+ image. Areas of green represent cropped land; brown represents summer fallow.

3. RESULTS AND DISCUSSION

Results from this study demonstrated that the RF classification algorithm applied to field-based image objects can provide high class accuracies in the discrimination of cropland from CR and crop from fallow (Table 1). An RF O-O classification based on May TM data was able to successfully separate CR from cropland with producer's accuracies of 90% and 100%, respectively. Previous pixel-based studies had relied on more elaborate multi-year change techniques to achieve similar accuracies (Egbert et al., 1998; Price et al., 1997). Classification error primarily resulted from the misclassification of CR as NT-cropped and tilled-crop. The misclassified sites were often those under recent conversion from cropland to CR, as was determined by an examination of data supplied through the Montana Farm Service Agency.

The ability to distinguish senesced crop from fallow with greater than 82% accuracy is considered to be highly acceptable, especially given the ability of the O-O-based RF model to separate stubble-laden fallow fields from those recently harvested. The RF variable importance plot indicated that object textural measures such as within-object contrast and homogeneity were often used as model predictive parameters, suggesting that object-derived information allowed for greater predictive ability under certain conditions.

Misclassification errors within the fallow category were attributed to objects located within landscapes characterized by narrow (< 100 m wide) crop and fallow strip management, due to within-pixel mixing of crop and fallow spectral signatures. The object-based classification tended to favor the "cropped" class, resulting in a classification bias under these conditions.

Model	Class	Producer's	User's
No-Till & Tillage			
May Pixel-based	NT	92%	79%
	Till	23%	48%
May Object-based	NT	91%	71%
	Till	31%	64%
Crop & CR			
May Object-based	Crop	100%	96%
	CR	90%	100%
Crop & Fallow			
August Object-based	Crop	95%	94%
	Fallow	82%	85%

Table 1. Classification (OOB) accuracy for tillage, CR, and crop status.

An object-based RF classification was not able to adequately distinguish tillage from NT (31% producer's accuracy), although accuracies were generally higher than was achieved with a pixel-based approach using these data. It was expected that RF, used in conjunction with an object-based approach, would have produced higher classification accuracies than those generated through a pixel-based, logistic regression approach. The O-O and pixel-based classifications, however, produced results very similar to those previously reported (Bricklemyer et al., 2006). The failure of RF to increase tillage class accuracy is likely due to a greater degree of spectral variability within management data utilized within this study, resulting from a larger number of study locations taken over a larger spatial area compared to the previous study. General difficulty in distinguishing tillage from NT has been reported in situations where the soil surface is covered by established crop canopy and plant residues (Daughtry et al., 2006; Gowda et al., 2001). An evaluation of RF model error showed that tilled-cropped locations were often misclassified as cropped NT or fallow NT, resulting in low producer's accuracies for both the RF-based O-O and pixel-based tillage classifications. We believe that similarities in surface residue cover between NT and tilled sites greatly attributed to the misclassification problem.

4. CONCLUSION

This study successfully applied the RF classification algorithm to spectral, textural, and neighborhood object-based parameters. Results from this study suggest that RF, used in conjunction with object-based data derived from moderate-resolution Landsat imagery, can accurately classify crop from CR and crop from fallow despite variability in the spectral data set. The incorporation of O-O methodology with RF efficiently allows for the integration of complex machine learning techniques with an advanced approach to image analysis. It also was found that classification predictions generated through randomForest could be easily incorporated into object database format for GIS-based spatial analysis.

REFERENCES

Baker, C., R. Lawrence, C. Montagne, and D. Patten.2006. Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models. *Wetlands*, 26: 465-474.

Benz, U.C., P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen. 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of Photogrammetry and Remote Sensing, 58: 239-258.

Breiman, L., and A. Cutler.2005. "Random Forests". http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (accessed 20 April 2008).

Breiman, L.2001. *RFs*. Machine Learning, 45: 5-32.

Breiman, L., J. Friedman, C. J. Stone, and R.A. Olshen.1998. *Classification and Regression Trees* (3ed Ed.). CRC Press, Boca Raton, Fl. 372 pp.

Bricklemyer, R.S., R.L. Lawrence, P.R. Miller, and N. Battogtokh.2006. Predicting tillage practices and agricultural soil disturbance in north central Montana with Landsat imagery. *Agriculture, Ecosystems and Environment*, 114: 210-216.

Chander, G., B. L. Markham, and J.A. Barsi.2007. Revised Landsat 5 Thematic Mapper radiometric calibration. *IEEE Geoscience and Remote Sensing Letters*, 4: 490-494.

Chubey, M.S., S.E. Franklin, and M.A. Wulder.2006. Object-based analysis of Ikonos-2 imagery for extraction of forest inventory parameters. *Photogrammetric Engineering and Remote Sensing*, 72: 383-394.

Crist, E.P., R. Laurin, and R.C. Cicone.1986. Vegetation and soils information contained in transformed Thematic Mapper data. In Proceedings of IGARSS '86 Symposium, Zurich, Switzerland, 8-11 September 1986. pp.1465-1470.

Daughtry, C.S.T., P.C. Doraiswamy, E.R. Hunt, Jr., A.J. Stern, J.E. McMurtrey III, and J.H. Prueger. 2006. Remote sensing of crop residue cover and soil tillage intensity. *Soil and Tillage Research*, 91: 101-108.

Egbert, S.L., Lee, R.Y., Price, K.P., and R. Boyce.1998. Mapping conservation reserve program (CRP) grasslands using multi-seasonal Thematic Mapper imagery. *Geocarto International*, 13: 17-24.

Fawcett, R., and D. Towery.2002. Conservation tillage and plant biotechnology: how new technologies can improve the environment by reducing the need to plow. *Conservation Technology Information Center*. 24 pp.

Gislason, P.O., J.A. Benediktsson, and J.R. Sveinsson.2006. RFs for land cover classification. *Pattern Recognition Letters*, 27: 294-300.

Gowda, P.H., B.J. Dalzell, D.J. Mulla, and F. Kollman.2001. Mapping tillage practices with Landsat Thematic Mapper based logistic regression models. *Journal of Soil and Water Conservation*, 56: 91-96.

Green, K. and C. Lopez.2007. Using object-oriented classification of ADS40 data to map the Benthic habitats of the state of Texas. *Journal of Photogrammetric Engineering and Remote Sensing*, 73: 861-865.

Ham, J., Y. Chen, M.M. Crawford, and J. Ghosh.2005. Investigation of the random forest framework for classification of hyperspectral data. *International Journal for Remote Sensing*, 243: 492-500

Hay, G.J., T. Blaschke, D.J. Marceau, and A. Bouchard.2003. A comparison of three image-object methods for the multiscale analysis of landscape structure. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57: 327-345.

Huang, C., B. Wylie, L. Yang, C. Homer, and G. Zylstra.2002. Derivation of a tasseled cap transformation based on Landsat 7 at-satellite reflectance. *International Journal of Remote Sensing*, 23: 1741-1748.

Lawrence, R.L., S. D. Wood, and R.L. Sheley.2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). *Remote Sensing of Environment*, 100: 356-362.

Lawrence, R., A. Bunn, S. Powell, and M. Zambon.2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90: 331-336.

Lawrence, R.L., and A. Wright.2001. Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote sensing*, 67: 1137-1142.

Navulur, K.2007. *Multispectral Image Analysis Using the Object-oriented Paradigm*. CRC Press, Boca Raton, FL.165 pp.

Pal, M.2005. RFs for land cover classification. *International Journal for Remote Sensing*, 26: 217-222.

Price, K.P., Egbert, S.L., Nellis, M.D., Lee, R.Y., and Boyce, R.1997. Mapping land cover in a High Plains agro-ecosystem using a multi-date Landsat Thematic Mapper modeling approach. *Transactions of the Kansas Academy of Science*, 100: 21-33.

Science Data Users Handbook,-Landsat 7 (SDH-L7).2006. National Aeronautics and Space Administration. http://landsathandbook.gsfc.nasa.gov/handbook/handbook_toc.html (accessed 10 March 2008).

Tucker, C.J. and P.J. Sellers.1986. Satellite remote sensing of primary production. *International Journal of Remote Sensing*, 7: 1395-1416.

Van Coillie, F. M.B., L. P.C. Verbeke, and R. R. De Wulf.2007. Feature selection by genetic algorithms in object-based classification of IKONOS imagery for forest mapping in Flanders, Belgium. *Remote Sensing of Environment*, 110: 476-487.

Yu, Q., P. Gong, N. Clinton, G. Biging, M. Kelly, and D. Schirokauer.2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogrammetric Engineering and Remote Sensing*, 72: 799-811.