

A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data

M. A. FRIEDL*, C. WOODCOCK, S. GOPAL, D. MUCHONEY,
A. H. STRAHLER and C. BARKER-SCHAAF

Department of Geography and Center for Remote Sensing, Boston University,
675 Commonwealth Avenue, Boston, MA 02215, USA

(Received 19 November 1998; in final form 18 March 1999)

Abstract. We present results from analyses conducted to evaluate the performance of advanced supervised classification algorithms (decision trees and neural nets) applied to AVHRR data to map regional land cover in Central America. Our results indicate that the sampling procedure used to stratify ground data into train and test sub-populations can substantially bias accuracy assessment results. In particular, we found spatial autocorrelation in test data to inflate estimates of classification accuracy by up to 50 points. Results from evaluations performed using independent train and test data suggest that the feature space provided by AVHRR NDVI data is poorly suited for most land cover mapping problems, with the exception of those involving highly generalized classes.

1. Introduction

A variety of efforts are currently underway to map land cover at continental and global scales using data from the Advanced Very High Resolution Radiometer (AVHRR) (e.g. Loveland and Belward 1997, DeFries *et al.* 1998, Friedl *et al.* 1999). However, rigorous assessment of the accuracy of these maps is proving difficult. This Letter discusses results from recent efforts conducted in preparation for the production of global land cover maps from the Moderate Resolution Imaging Spectro-Radiometer (MODIS). As part of our efforts we have noted several important aspects of the accuracy assessment process that we believe merit wider attention in the remote sensing and land cover community. Specifically, our work has demonstrated important limitations to commonly used accuracy assessment procedures that lead to the assignment of spuriously high accuracy to map products. Further, our results suggest serious limitations of AVHRR data for the purpose of land cover mapping.

2. Methods

2.1. Data and algorithms

The results reported here are derived from analyses conducted using a regional data set from Central America (Muchoney *et al.* 1999). The objectives of these analyses are to test and refine supervised classification algorithms being developed

* e-mail: friedl@bu.edu

for use with MODIS data. For full details regarding the MODIS land cover product see Strahler *et al.* (1996).

Our algorithm development effort is directed toward two relatively new supervised classification models: the Fuzzy ARTMAP artificial neural network (Carpenter *et al.* 1992), and C5.0 (Quinlan 1993), a univariate decision tree. These algorithms will use MODIS data to map global land cover at 1 km spatial resolution on a quarterly basis using the classification scheme defined by the International Geosphere-Biosphere Programme (IGBP) (Belward 1996). This classification scheme includes 17 broad climate-independent classes of land cover designed for use within environmental models. In the absence of MODIS data, we have been using AVHRR data to test and refine our algorithms and data processing flow.

The Central America dataset used for this work is composed of two main elements (Muchoney *et al.* 1999). The first is a time series of co-registered normalized difference vegetation index (NDVI) data produced from monthly composited AVHRR data covering the period from April of 1992 to March of 1993 (i.e. 12 images total). The land area in this data set encompasses 619 048 km² and includes the countries of Belize, Guatemala, El Salvador, Nicaragua, Costa Rica and Panama. The second element includes 428 sites within Central America where IGBP land cover labels have been assigned based on manual interpretation of Landsat Thematic Mapper (TM) data. For details regarding the specifics of the site sampling and data generation please see Muchoney *et al.* (1999). These sites were used to both train and test C5.0 and Fuzzy ARTMAP using mutually exclusive subsets of the data. As a basis for comparison, we conducted parallel tests using two more conventional supervised classification algorithms: maximum likelihood and 1-nearest neighbour (1-NN).

2.2. Accuracy assessment procedure

To assess the classification accuracy produced by each of the supervised algorithms identified above we conducted five randomized trials using mutually exclusive splits of the site data. For each trial 80% of the data were used to train each classification algorithm, while 20% were held out and used to evaluate classification accuracy based on data that were not 'seen' in the training phase. These randomized trials were performed in two different ways. First, the data were stratified into independent train and test sets using pixels selected at random from the entire dataset, independent of the site from which they were derived. Second, the data were randomly stratified into independent train and test groups keeping data from individual sites together. We will refer to the former method as 'pixel-based splits' and the latter as 'site-based splits'. Our results therefore include 40 distinct train and test classification runs: four classification algorithms, five 80/20 splits, and two different splitting procedures.

3. Results

Table 1 presents average results across the five train and test splits for each algorithm and for each splitting procedure. Several results are obvious from this table. First, for the pixel-based splits, all three of the nonparametric classification algorithms (C5.0, Fuzzy ARTMAP, and 1-NN) provide much higher accuracies than does the more conventional and parametric maximum likelihood algorithm. This is consistent with previous results and supports our use of more flexible and distribution-independent algorithms such as artificial neural networks and decision trees. For classifications generated using site-based splits, accuracies range from 16–54

Table 1. Summary of classification accuracies across five cross-validation runs for each classification algorithm using pixel and site-based splits.

Algorithm	Pixel-Based Accuracy (%)	Site-Based Accuracy (%)
Maximum Likelihood	52.4	36.6
1-Nearest Neighbour	91.5	37.9
Fuzzy ARTMAP	79.3	51.4
C5.0	88.9	42.6

percentage points lower than those generated using pixel-based splits. Further, the decision tree and 1-NN algorithms produce accuracies that are only marginally better than those produced by the maximum likelihood algorithm, and the highest accuracy from all four algorithms was only 51.4%.

Consideration of the results from the 1-NN algorithm provides the simplest interpretation for the results observed across all three nonparametric algorithms. Specifically, these results show that for 91.5% of the test cases there exists one training example of the same class for which the Euclidian distance between that training case and the test case is minimum. The mechanism producing this result is revealed via graphical inspection of the data. Specifically, figure 1 plots the NDVI test data for two classes (grassland and evergreen needleleaf forest) using the algorithm described by Friedman and Rafsky (1981) to visualize multivariate data. This algorithm uses minimal spanning trees to compute the multivariate distance between points and where the projection onto the plane preserves the distances between points. In this figure, the main variability in the 12-month NDVI data is therefore collapsed into two dimensions and the data from individual sites have been assigned distinct symbols (note that different sites have different numbers of samples). This plot shows that strong clustering is present in the NDVI data within sites and that therefore the NDVI data possess substantial spatial autocorrelation at the scale of

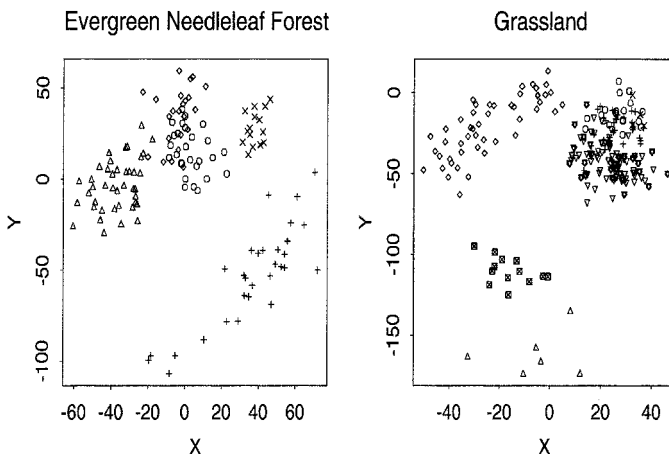


Figure 1. Two-dimensional presentation of the 12-month NDVI test data for the evergreen needleleaf forest and grassland classes. Data from individual sites are given distinct symbols in each plot (five and seven sites, respectively).

individual sites (e.g. Belward and Lambin 1990). A by-product of this spatial autocorrelation is that the 1-NN algorithm is very effective at labelling unseen data because each test case is correlated with training cases extracted from the same site.

The specifics that give rise to similar patterns in the results produced by C5.0 and Fuzzy ARTMAP are more complex, but the underlying mechanisms are the same. Like the 1-NN algorithm, the neural network and decision tree algorithms both use many-to-one mapping to assign class values to observations. Further, these algorithms are adept at detecting subtle patterns present in training data. As a result, both algorithms are effective at detecting clustering in the data caused by spatial autocorrelation at the site level. In other words, C5.0 and Fuzzy ARTMAP are 'learning' the patterns associated with each of the sites, rather than patterns associated with the individual classes. Therefore, both C5.0 and Fuzzy ARTMAP produce relatively high accuracies for pixel-based splits. In contrast, when site-based splits are used, the classification accuracies decline precipitously.

4. Conclusions

The results described above point to two main conclusions. First, because spatial autocorrelation is present in the AVHRR data at the site level, pixel-based splits do not provide independent train and test data, and therefore do not provide a useful basis for evaluating map accuracies. Second, the feature space provided by the 12-month time series of NDVI data is inadequate to predict the classes defined by the IGBP classification in Central America. While the results presented in this Letter are specific to the Central America dataset that we examined, it is probable that these conclusions are relevant to the general problem of using AVHRR data to map IGBP-like classes of land cover.

As a working hypothesis, we ascribe the high autocorrelation among pixels and the low correlation among sites in AVHRR data to two phenomena. First, with classes as broad as those of the IGBP, there are clearly unique subtypes that have different manifestations in the AVHRR dataset but are of the same IGBP class. The way to accommodate these subtypes is to ensure that all are sampled for training in a final classification run. In the present Central America dataset, we believe that nearly all such subtypes have been sampled. However, in some random splits, all examples of less common subtypes may be omitted from the training set, and thus will be misclassified in the test set due to lack of exemplars.

Second, preliminary analysis suggests that sites within subtypes that are similar on the ground appear dissimilar in the AVHRR data due to spatially autocorrelated noise or differences in vegetation seasonality. An obvious source of this noise is the maximum-value compositing procedure used to generate the monthly sets in association with the low signal-to-noise ratio of AVHRR data for land cover mapping applications. For example, one or more subregions containing training sites may be obscured by clouds during a particular month, creating a low NDVI and falsely inducing a unique temporal signature for those sites. We believe that this phenomenon is responsible for a large proportion of the test site misclassifications reported above.

Finally, it is important to note that the low site-based classification accuracies we observe here are not characteristic of single-date, high-resolution Landsat TM data, for which numerous studies have shown much higher classification accuracies. We believe that the absence of cloud contamination and greater spectral information content of Landsat TM data account for this difference. This observation offers

encouragement for success at land cover classification with MODIS data, as atmospheric correction and rigorous cloud clearing will improve the quality of individual observations while the band complement of MODIS will provide the spectral information lacking in AVHRR NDVI data. When this information is coupled with compositing based on data quality and suitability, rather than maximum value, these attributes should provide a superior information base for land cover classification with MODIS.

Acknowledgments

This work was supported by NASA Grants NAG5-7218 and NAS5-31369. The hard work of the land cover 'troops' at BU who helped to assemble the site database is gratefully acknowledged as are the staff of the Global Land 1-km AVHRR project at Eros Data Center who kindly provided the 1-km AVHRR data.

References

- BELWARD, A., 1996, The IGBP-DIS Global 1 km Land Cover Data Set DISCOVER: Proposal and Implementation Plans. IGBP-DIS Working Paper, IGBP-DIS Office, Meteo—France, 42 Av. G. Coriolis, F—31057, Toulouse, France.
- BELWARD, A., and LAMBIN, E., 1990, Limitations to the identification of spatial structures from AVHRR data. *International Journal of Remote Sensing*, **11**, 921–927.
- CARPENTER, G., GROSSBERG, S., MARKUZON, N., REYNOLDS, J., and ROSEN, D., 1992, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, **3**, 698–713.
- DEFRIES, R., HANSEN, M., TOWNSHEND, J., and SOHLBERG, R., 1998, Global land cover classifications at 8 km spatial resolution: The use of training data derived from Landsat Imagery in decision tree classifiers. *International Journal of Remote Sensing*, **19**, 3141–3168.
- FRIEDL, M. A., BRODLEY, C., and STRAHLER, A., 1999, Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 969–977.
- FRIEDMAN, J., and RAFSKY, L., 1981, Graphics for the multivariate two-sample problem (with discussion). *Journal of the American Statistical Association*, **76**, 277–295.
- LOVELAND, T., and BELWARD, A., 1997, The IGBP-DIS global 1 km land cover data set, DISCOVER: first results. *International Journal of Remote Sensing*, **18**, 3289–3295.
- MUCHONEY, D., BORAK, J., CHI, H., FRIEDL, M., GOPAL, S., HODGES, J., and STRAHLER, A., 1999, Application of the MODIS global supervised classification model to vegetation and land cover mapping of Central America. *International Journal of Remote Sensing*, in press.
- QUINLAN, J., 1993, *C4.5: Programs for Machine Learning* (San Mateo, CA: Morgan Kaufmann).
- STRAHLER, A., TOWNSHEND, J., MUCHONEY, D., BORAK, J., FRIEDL, M., GOPAL, S., HYMAN, A., MOODY, A., and LAMBIN, E., 1996, *MODIS Land Cover Product Algorithm Theoretical Basis Document (ATBD), V4.1* (Boston: Boston University Center for Remote Sensing).