A METHOD FOR SELECTING TRAINING DATA AND ITS EFFECT ON AUTOMATED LAND COVER MAPPING OF LARGE AREAS

René R. Colditz*, Michael Schmidb, Rainer Ressl, Matthew C. Hansen, Stefan Dech

a National Commission for the Knowledge and Use of Biodiversity (CONABIO), Avenida Liga Periférico – Insurgentes Sur No 4903, Col. Parques del Pedregal, Del. Tlalpan 14010 México, D.F.
b German Aerospace Center – German Remote Sensing Data Center (DLR-DFD), 82234 Wessling Germany
c Geographic Information Science Center of Excellence (GIScCE) – South Dakota State University, 1021 Medary Ave., Wecota Hall Box 506B, Brookings, SD 57007, USA
d Department of Geography – University of Wuerzburg, Am Hubland, 97074 Wuerzburg, Germany

* corresponding author rcolditz@conabio.gob.mx (Tel: +52-55-5004-5020, FAX: +52-55-5004-4931)

ABSTRACT

Many remote sensing projects require the utilization of sample data for training a supervised classification algorithm. In almost all instances the correctness of training data is highly important for accurate image classification. While data obtained during field studies are suitable for many small scale studies and the classification of high and very high spatial resolution data, automated procedures are necessary to map large areas with medium to coarse resolution images. Most coarse resolution land cover classifications are based on previous studies and mapping efforts.

This study illustrates a procedure for training data selection of coarse resolution images if a high resolution map already exists. In a second step the paper analyzes the significance of training data selection parameters for classification accuracy.

Figure 1: Simplified examples of homogeneity measure.

Assuming a perfect geometric alignment between the high and coarse resolution data, the proportion of every class in each coarse resolution cell is calculated. In essence, this algorithm computes a fractional class estimation of the coarse cell based on a high spatial resolution map. This result is denoted as “homogeneity” and exemplarily illustrated in Figure 1. While Examples A and B indicate mixed coarse resolution cells with similar proportions of all classes, examples C and D are comparatively homogeneous with a clear majority belonging to one class. Pixels suitable as training data should belong to one land cover type; if not solely then at least mainly. Similar approaches have been successfully employed by [1, 2, 3].

However, a perfect overlay between datasets is commonly impossible. A more conservative estimate can be achieved if the area of analysis is enlarged into neighboring cells (see Figure 2). On the one hand this
procedure limits the sample candidates. On the other hand, the confidence in the selection of training pixels increases. Due to the inverse relation between available samples and higher certainty with increased neighborhoods, there is a critical cut-off value because a minimum number of samples should be available to draw a representative training data set. A very strict limitation of sampling may not sufficiently characterize multimodal frequency distributions of all classes in the feature space, resulting in lower classification accuracies.

Figure 2: Simplified example of homogeneity for extended neighborhoods.

The study is based on MODIS time series metrics and employs a set of decision trees. This classifier, similar to random forest, derives a fuzzy image classification. Tests in Germany and South Africa yielded on average a 10% increase in classification accuracy. Furthermore, the presented up-scaling algorithm results in a fuzzy reference map which can be employed for assessment of a fractional image classification; a difficult and less explored field in remote sensing [4]. Therefore this study presents both, conventional error matrix statistics as well as fuzzy error estimates to illustrate the importance for using high certainty training pixels.

KEYWORDS
Training Data Selection, Image Classification, Decision Tree Classifier, MODIS, Accuracy Assessment

REFERENCES