AN AUTOMATED ALGORITHM TO SCREEN MASSIVE TRAINING SAMPLES FOR A GLOBAL IMPERVIOUS SURFACE CLASSIFICATION

TAN BIN[1,2], COLSTOUN BROWN DE ERIC[1], WOLFE E. ROBERT[1], TILTON C. JAMES[1], HUANG CHENGQUAN[3], SMITH E. SARAH[1]

[1]NASA GSFC, Greenbelt, USA.
[2]ERT, Laurel, USA.
[3]Department of Geography, UMD, Maryland, USA.

Bin.Tan@nasa.gov

An algorithm is developed to automatically screen the outliers from massive training samples for Global Land Survey – Imperviousness Mapping Project (GLS-IMP). GLS-IMP is to produce a global 30 m spatial resolution impervious cover data set for years 2000 and 2010 based on the Landsat Global Land Survey (GLS) data set. This unprecedented high resolution impervious cover data set is not only significant to the urbanization studies but also desired by the global carbon, hydrology, and energy balance researches. A supervised classification method, regression tree, is applied in this project. A set of accurate training samples is the key to the supervised classifications. Here we developed the global scale training samples from 1 m or so resolution fine resolution satellite data (Quickbird and Worldview2), and then aggregate the fine resolution impervious cover map to 30 m resolution. In order to improve the classification accuracy, the training samples should be screened before used to train the regression tree. It is impossible to manually screen 30 m resolution training samples collected globally. For example, in Europe only, there are 174 training sites. The size of the sites ranges from 4.5 km by 4.5 km to 8.1 km by 3.6 km. The amount training samples are over six millions. Therefore, we develop this automated statistic based algorithm to screen the training samples in two levels: site and scene level. At the site level, all the training samples are divided to 10 groups according to the percentage of the impervious surface within a sample pixel. The samples following in each 10% forms one group. For each group, both univariate and multivariate outliers are detected and removed. Then the screen process escalates to the scene level. A similar screen process but with a looser threshold is applied on the scene level considering the possible variance due to the site difference. We do not perform the screen process across the scenes because the scenes might vary due to the phenology, solar-view geometry, and atmospheric condition etc. factors but not actual landcover difference. Finally, we will compare the classification results from screened and unscreened training samples to assess the improvement achieved by cleaning up the training samples.

Keywords: GLS-IMP, Landsat, Regression tree, Spectral space analysis