Applying Unsupervised Classification to Map Guatemala City

Author: David Kolodziejski
Date: July 19th, 2021

Introduction

Guatemala City is a mountainous city in Central America surrounded by forests, grasslands, and cropland. This city is the capital of Guatemala and is home to approximately 995,130 according to macrotrends.net. On November 9th, 2008 the ASTER Satellite captured an image of Guatemala City with a 15-meter resolution and is composed of the green band, red band, Near Infrared (NIR) band, SWIR 1 band, SWIR 2 band, and thermal band. For this analysis, a land classification map was created through supervised classification methods through the ENVI environment. The user identified the various landcover classifications by delineating Regions of Interest (ROI) to supplement the supervised classification. Two separate classification maps were generated using the Maximum Likelihood Classification method and the Supported Vector Machine algorithm. Following the creation of these maps, an accuracy assessment was conducted by classifying randomized test sites (ground truth pixels) and cross-validating those classifications with our final product to create a confusion matrix highlighting user accuracy and producer accuracy.

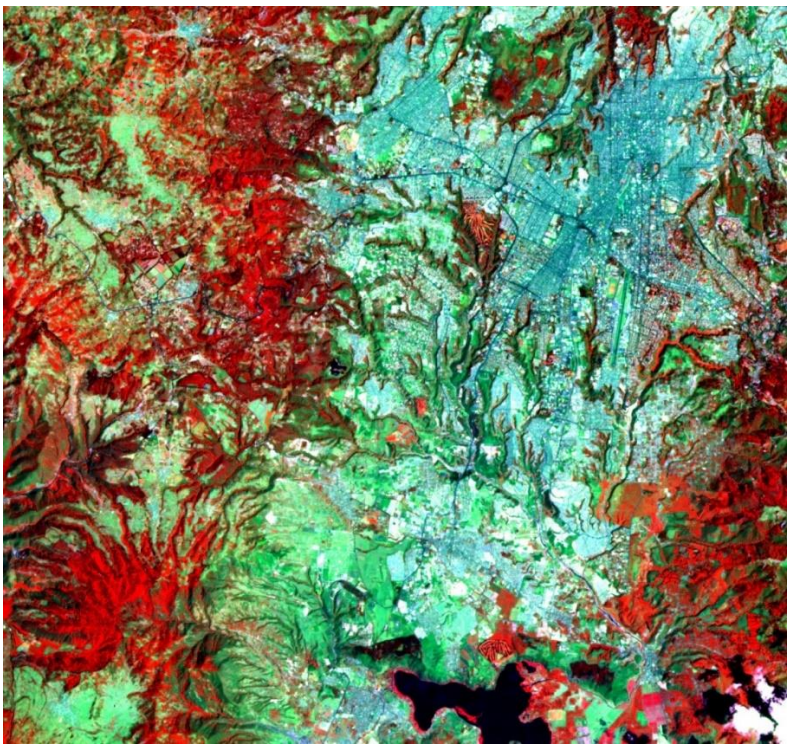

Figure 1: False Color image of Guatemala City taken from the ASTER Satellite on November 9th, 2008: Bands are set to the following RGB Bands of the monitor: NIR to Red, Red to Green, Green to Blue

Methods

For this analysis, two land classification maps were generated using the Maximum Likelihood and Supported Vector Machine supervised classification. The supervised classification process involves using regions of interest (training data) to create a thematic map. Training data is created by the user to delineate examples of the various classifications that are desired to be displayed. Each region of interest (ROI) contains the spectral signatures of the pixels across all bands that fall underneath it. When the supervised classification is run, the pixels that were not included within our ROIs get compared to the spectral signatures contained in each training sample and are then assigned automatically.

Maximum Likelihood

The Maximum Likelihood Classifier compares the spectral signatures within our ROIs to the remaining unknown spectral signatures and determines how pixels are classified based on the highest probability given the spectral signals in our training data. This classification method determines the means, variance, and convergence of the training data and compares these metrics with the unknown classification to produce a thematic map. The advantage of a supervised algorithm Is that is it the most accurate of the classifiers if the spectral signatures are normally distributed. For example, if an ROI has spectral signatures that are intended to represent classifications (e.g. forest vs urban) that the output may not be completely accurate. Therefore, one of the many disadvantages of this method is that is heavily reliant on the user to accurately and carefully generate training area. If training areas are not collected appropriately, then the classification will produce poor results. An additional disadvantage is the classification is non-spatial and does not identify patterns. As a result, individual pixels will be present in large classification and maybe too descriptive for our analysis. Conversely, this may be desired by the user and may act as a benefit. Running the Maximum Likely Classification is computationally intense and will require a significant amount of time to process.

Supported Vector Machines

In addition to producing a map using the Maximum Likelihood supervised classification, An alternative method to producing a thematic map through supervised classification is Supported Vector Machine (SVM) Learning. This machine-learning algorithm uses the spectral signatures within our ROIs and determines the optimal hyperplane. The optimal hyperplane is a decision boundary the algorithm uses to determine how pixel values should be separated (i.e. classified) and is the maximum distance away from the nearest separated pixels. The pixel values that are closest to the optimal hyperplane are the support vectors. This process is then repeated until all pixels have been evaluated separated/classified by the optimal hyperplane. An enormous advantage of using SVM is that can handle non-linear boundaries through the use of kernel functions. Kernel functions allow us to transform the data into a "3D Space" that allows us to perform the linear class separation. This is incredibly beneficial as it allows us to be more detailed in our overall classification. Additionally, SVM allows us the flexibility to separate classes using two approaches with our ROI data: one class vs. all classes (forest vs. everything else) or one class vs. one class for all class pairs (forest vs wetland, forest vs cropland, etc.). Both parameters will likely produce good results but the latter option will take additional time to process. Thus, one of the disadvantages of the SVM is that it can take an enormous amount of time to process depending on the number of classes you have; more than seven classes will cause processing speeds to diminish. Furthermore, if the outputted results are not what you were expected you will likely need to delete ROIs instead of adding additional ones. This is because the SVM algorithm is highly dependent on the pixel

values that are closest to the optimal hyperplane (support vectors). As a result, the user will likely have to recreate ROIs to significantly alter the outputted results.

Accuracy Assessment

An accuracy assessment was conducted to determine the validity of the thematic maps produced through the Maximum Likelihood and Supported Vector Machine supervised classification. This assessment utilized randomized test sites (i.e. pixels) for the user to correctly identify to then determine how accurate landcover classifications are in both of our maps (Figure 02). Test sites are randomized to limit/remove bias influence from the assessment. For this analysis, 52 test sites were randomly selected to be classified and compared without results. This unbiased approach allows the user to properly assess the quality of their map and address any deficiencies that may be present.

If the overall or individual land cover classification accuracies are low, then the user can properly address investigate these areas before moving forward with the analysis. For example, If the land cover classification accuracy is low, then it is recommended that the user evaluate the ROIs and test sites to determine if any user-related deficiencies came from digitizing. Oftentimes, ROIs will unintentionally include spectral signatures from various other classes that will affect the overall accuracy assessment. If the ROIs are properly delineated, then the next step would be to review the test site. Oftentimes, the most difficult pixels to classify include instances where the randomized pixel falls on in transitional areas (Figure 02). If these areas are classified incorrectly, then the producer's accuracy will drop.

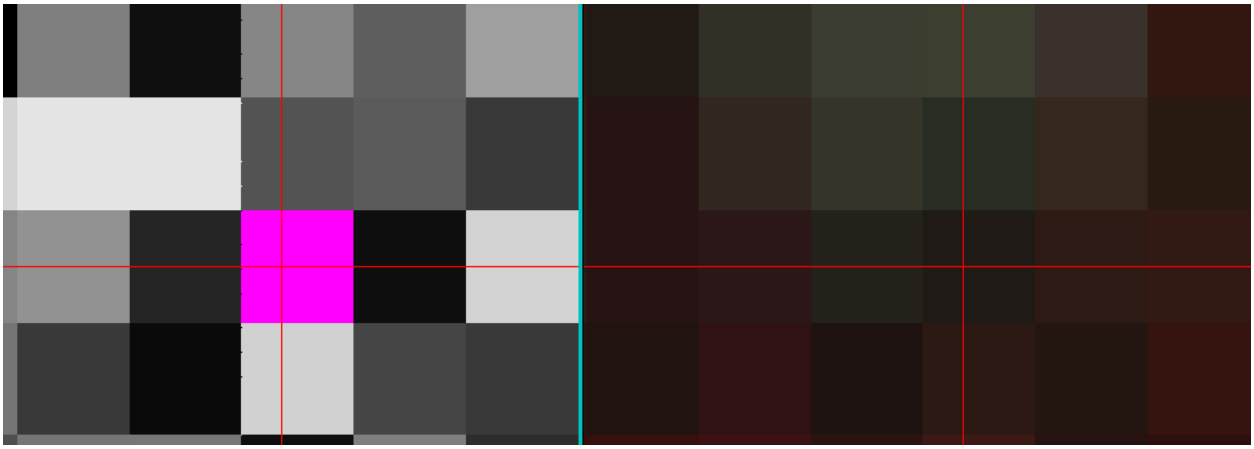

Figure 2: Classifying randomized pixels to be used from Accuracy Assessment. The pink pixel (left) is a randomized pixel generated by ENVI and the right image show the False Color image that corresponds to that pixel.

Results

The Land Classification of Guatemala City generated from the Maximum Likelihood and Supported Machine Vector classifications are illustrated below. These maps attempt to identify Forest, Grassland, Water, Harvested Cropland, Lush Cropland, Algae, Cloud Cover, Urbanization, and Undistinguishable classifications. The accuracy assessment indicated that the Support Vector Machine map is the more accurate product with a 69.23% accuracy. Conversely, the accuracy assessment indicates that the grassland classification in both the Maximum Likelihood and Supported Vector Machine are inaccurate with an accuracy of 14.29%.

The most distinguishable difference between the Maximum Likelihood (Figure 03) and Supported Vector Machine (Figure 04) Maps and the class representation of Forest and Grassland. The Maximum Likelihood map displays significantly more grassland than the support vector machine output. This is most noticeable in the western half of the image. Urbanization, Water, and Cropland classifications are fairly similar across both supervised classifications.

NOTE: Cloud Cover, undistinguishable, Harvest Cropland, Lush Cropland, and Algae were not identified in our original test sites. Therefore, Cloud Cover, Undistinguishable, and Algae were excluded from our overall assessment. Harvest Cropland and Lush Cropland were intentionally added to our random sample to allow the accuracy assessment to run successfully and display how the producer's accuracy is influenced by these classifications.

Maximum Likelihood: Guatemala City Accuracy

The accuracy assessment indicated that the thematic map produced by the Maximum Likelihood supervised classification is approximately 49.09% (Table 01). The largest factors that influence the overall accuracy were between the Forest and Grassland Classification. Of the 52 pixels that were utilized for this assessment, our thematical map designated 14 pixels as Forest, while our test sites that 27 pixels as Forest. Thus, the thematic map is approximately 51.85% accurate (producer's accuracy) in comparison to the test sites. Conversely, the User's Accuracy for Forest is approximately 93.33% (Table 01). This indicates that too many pixels are classified as Forest in our Maximum Likelihood map than what our test sites would anticipate. The remaining 13 pixels that were omitted were primarily classified as grassland.

The grassland classification is one of the lowest User and Producer accuracies according to this assessment. The user's accuracy is 9.09% while the producer's accuracy is approximately 14.29% (Table 01). In general, classifying grassland posed the most challenges of all the landcover classification. This is because most of the grassland occurs in transitional areas and small isolated locations in heavily forest areas (Figure 01). As a result, ROIs and test sites need to be undeniably grassland or the thematic map will produce poor results and the accuracy assessment will indicate that. Of the 7 pixels identical to grassland in our test sites, only a single pixel matched with our map. The remaining pixels were classified as either forest or harvest cropland (Table 01).

Supported Vector Machine: Guatemala City Accuracy

The accuracy assessment indicated that the thematic map produced by the Supported Vector Machine supervised classification is 69.23% (Table 02) and approximately 20% more accurate than the map produced by the Maximum Likelihood supervised classification. Overall, the producer's accuracy increased indicated that the map now correlates better to what was expected in our test sites. In comparison to the Maximum Likelihood map, the producer's accuracy of the forest increased from 51.85% to 74.07% (Table 02) and urbanization increased from 61.54% to 76.92% (Table 02). However, the producer's accuracy for the grassland classification did not change between Maximum Likelihood and Supported Vector Machine supervised classifications.

Maximum Likelihood: Guatemala City

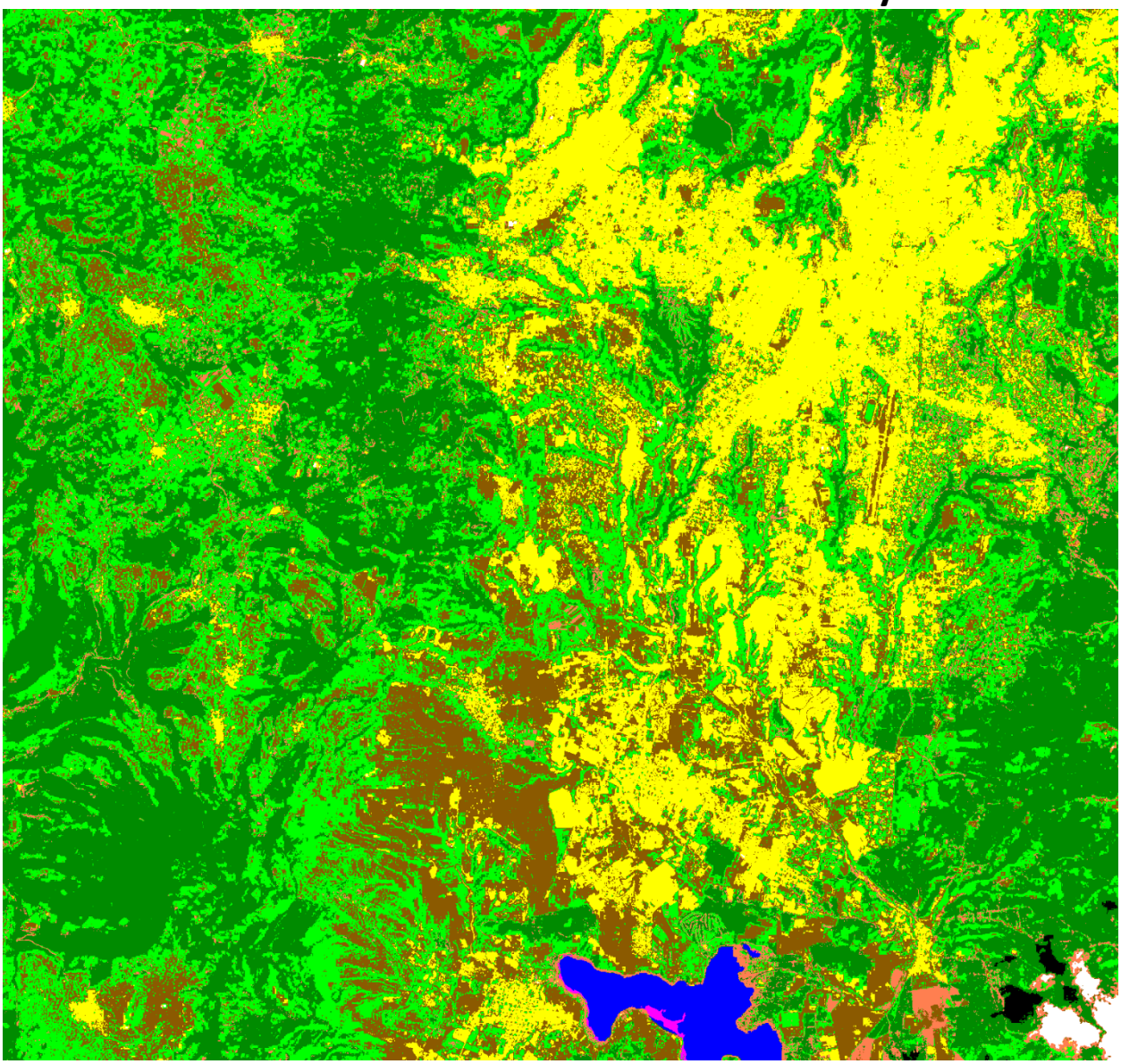

Figure 3: Maximum Likelihood Supervised Classification results of Guatemala City

Table 1: Accuracy Assessment of Maximum Likelihood Supervised Classification. Harvested Cropland and Lush Cropland were NOT created from a randomized test site and were manually included to complete the accuracy table.

Class	Water	Forest	Grassland	Harvested Cropland	Lush Cropland	Urbanization	Total	User's Accuracy
Water	3	0	0	0	0	0	3	100.00%
Forest	0	14	1	0	0	0	15	93.33%
Grassland	0	10	1	0	0	0	11	9.09%
Harvested Cropland	0	1	5	0	0	5	11	0.00%
Lush Cropland	0	2	0	0	1	0	3	33.33%
Urbanization	0	0	0	1	0	8	9	88.89%
Total	3	27	7	1	1	13	52	
Producer's Accuracy	100.00%	51.85%	14.29%	0.00%	100.00%	61.54%		49.09%

Supported Vector Machines: Guatemala City

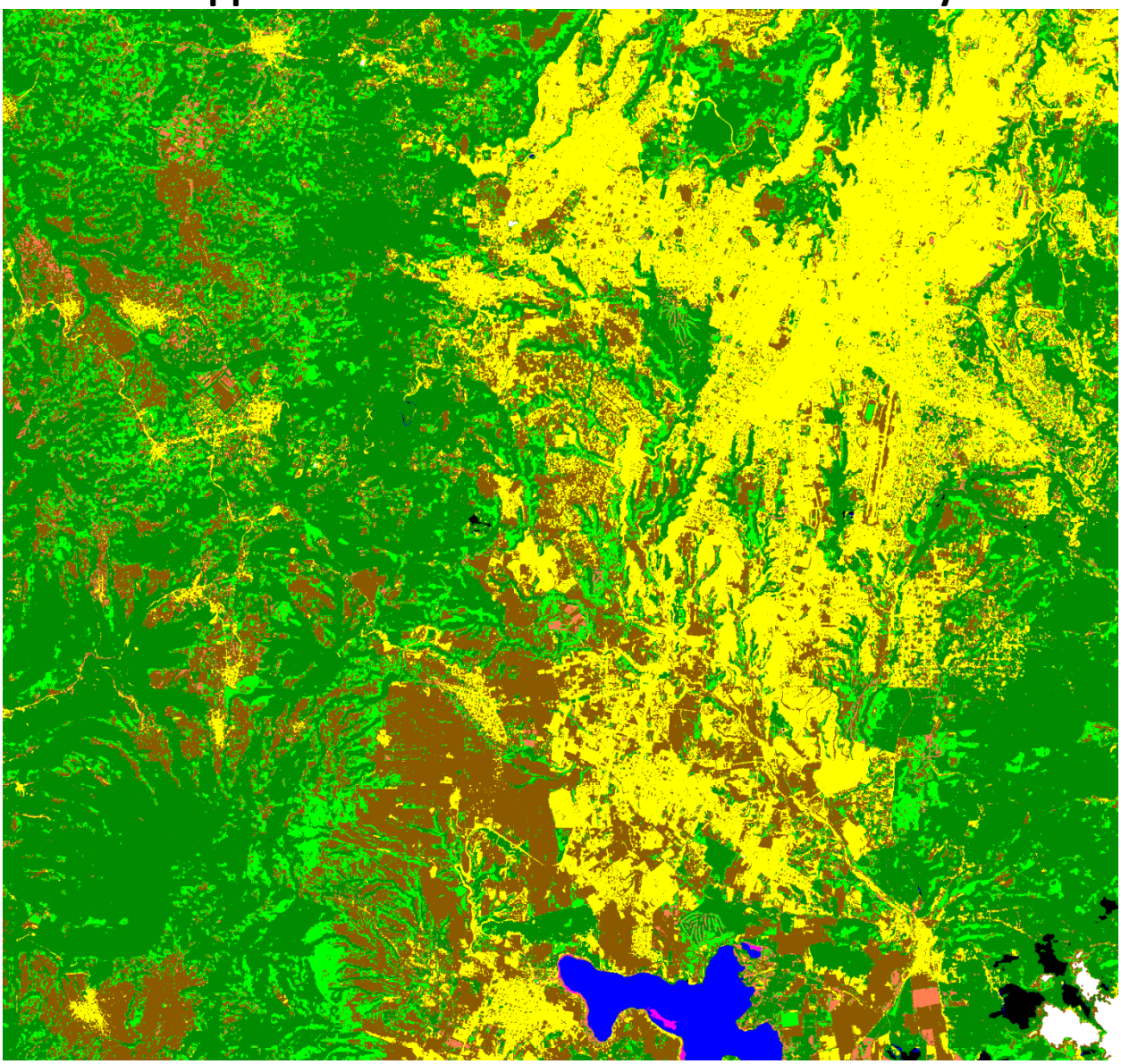

Figure 4: Supported Vector Machine Supervised Classification result of Guatemala City

Table 2: Accuracy Assessment of Support Vector Machine Supervised Classification. Harvested Cropland and Lush Cropland were NOT created from a randomized test site and were manually included to complete the accuracy table.

Class	Water	Forest	Grassland	Harvested Cropland	Lush Cropland	Urbanization	Total Pixels	User's Accuracy
Water	3	0	0	0	0	0	3	100.00%
Forest	0	20	1	0	0	0	21	95.24%
Grassland	0	3	1	0	0	0	4	25.00%
Harvested Cropland	0	2	5	1	0	3	11	9.09%
Lush Cropland	0	0	0	0	1	0	1	100.00%
Urbanization	0	2	0	0	0	10	12	83.33%
Total Pixels	3	27	7	1	1	13	52	
Producer's Accuracy	100.00%	74.07%	14.29%	100.00%	100.00%	76.92%		69.23%

Discussion

Throughout this analysis, the grassland land cover type was the most difficult to classify. When collecting training areas to produce the supervised classification, grassland and harvested cropland appeared very similar on imagery. This made it difficult at times to properly distinguish from the two classifications and likely led to accuracy errors in the map (as indicated by the accuracy assessment). Furthermore, grasslands occurred primarily in transitional areas (i.e.forest to grassland to urbanization) were the size of the classification was narrow. In future analysis, it would be best to distinguish transitional areas using the "point" option in the ROI tool.

Another limitation that became apparent with this analysis came from the accuracy assessment and producing test sites for User's accuracy. Of the 9 classes that were identified in our supervised classification, test sites only accounted for 4 classifications. Furthermore, of the 4 classifications, Forest made up approximately 52 percent of all test sites. Although the test sites were produced randomly, I do not feel that there was enough representation of other classes to properly assess the overall accuracy. For example, only 7 test sites were identified as grassland, and Harvested Cropland, Lush Cropland, and Algae were not represented. The test sites of these "missing" classifications were intentionally added to along ENVI to produce the accuracy assessment. However, this creates an inaccurate representation of the assessment where the classifications exhibited a nearly 100% accuracy rate. The accuracy table was then adjusted to exclude these variables to provide a better representation of the overall accuracy. For further analysis, increasing the number of test sites will likely account for all classifications and provide an adequate sample size.