GEE Classification Implementation Guide

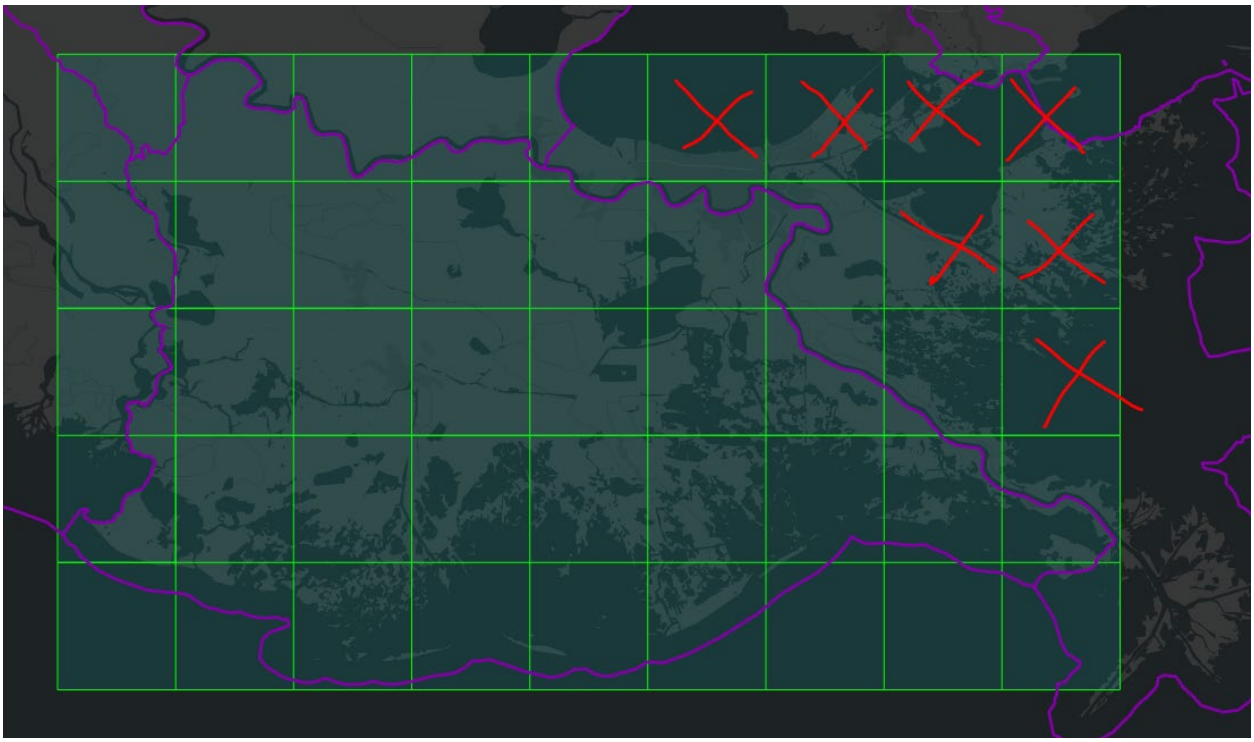
July 5th 2023

Courtney Di Vittorio

Example with Barataria – watershed #8

Step 1: Identify watershed sub-grids that should not be included in the GEE classification.

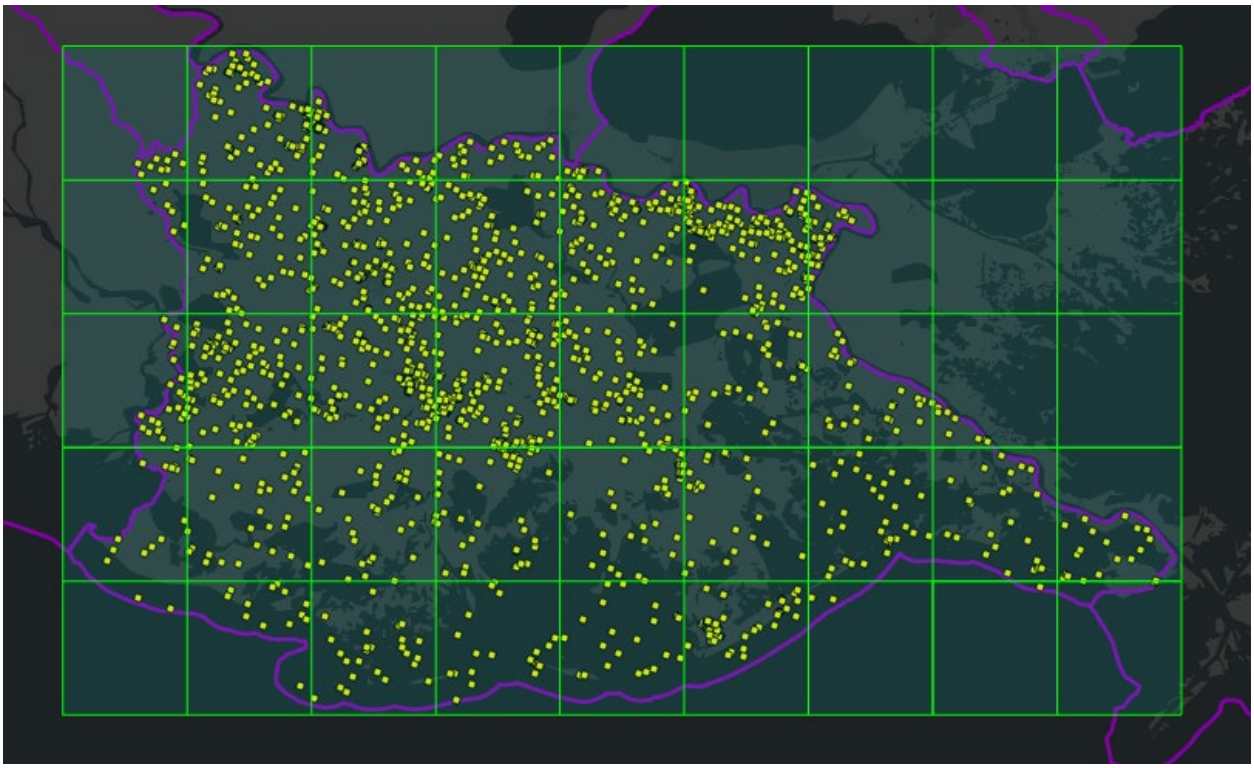
a) In ArcPro load the watershed extents and add the sub-grid kml files for your watershed (located in "Coastwide_Classification/watershed_grids/grid_kml". An example for Barataria (watershed #8) is shown below.
b) Identify the rows and columns that should be removed. In this case it is r1c6 (row 1 column 6), r1c7, r1c8, r1c9, r2c8, r2c9, r3c9.



Step 2: Identify the grids that do not have training/validation data for each of the 5 cross-validation datasets.

a) Under the map tab of ArcPro, click the drop-down arrow under add data and then click "XY Point Data". Select one of the training data csv files from "Coastwide_Classification\training_data\gee_repSites_allClasses". For the representative sites we will look at each cross validation set individually (trainingBARcv1.csv, trainingBARcv2.csv, …)
b) On the geoprocessing side bar, make sure Longitude is populated for the "X Field" and Latitude is populated for the "Y Field" (should be automatic). Click Run.
c) Similar to step 1, add the sub-grid kml files for your watershed to the map.

d) Identify grids that overlap with the watershed that do not have any training data. trainingBARcv1.csv is shown below, and there are no grids that meet this criteria, but it is likely that this criteria will be met for other datasets and other watersheds.

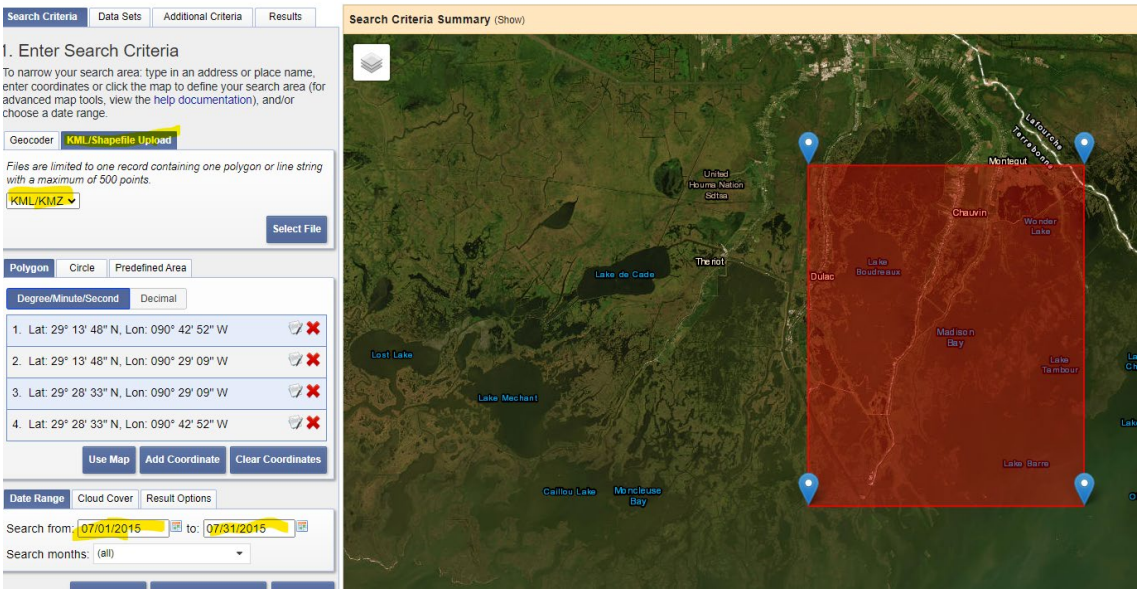


Step 3: Run GEE Segment export

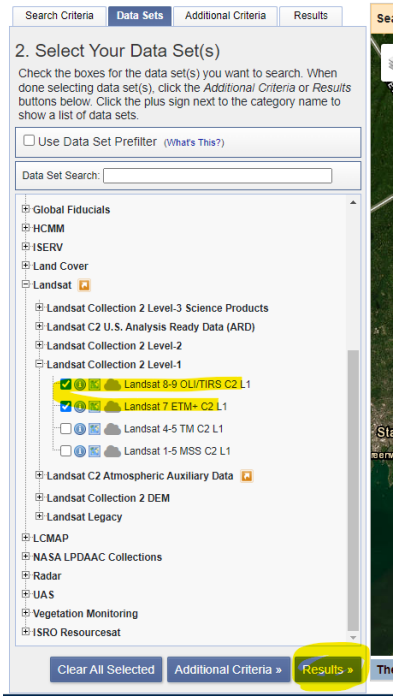
a) Open up GEE script ccdc_segments_template_step1
b) Add your name and date to the top and use “save as” to change the name to ccdc_segments_[watershed_number]_[your_initials]
c) Open txt file that contains lat and lon coordinates for your watershed
d) Copy/paste the grid coordinates for your watershed at the top
e) Delete any grids that are outside of the watershed boundaries
f) IF the segment does not export then this is likely because the grid is located in a place where there are 3 overlapping Landsat tiles, which creates a very dense stack of images with duplicate information and causes GEE to run out of memory. Go through Step 3 supplement for that grid to confirm whether issues is 3 overlapping tiles and identify the 2 tiles you would like to keep. These two tiles should have the same path or row number. Take note of the common number. Open the ccdc_segments_redMemory_template GEE script and save a copy, adding the watershed number and your initials to the end. Copy/paste the grid you are processing and enter the common path or row number. Export the results as an asset.

Step 3 (when running out of memory due to tile overlap): Evaluate overlapping Landsat tiles

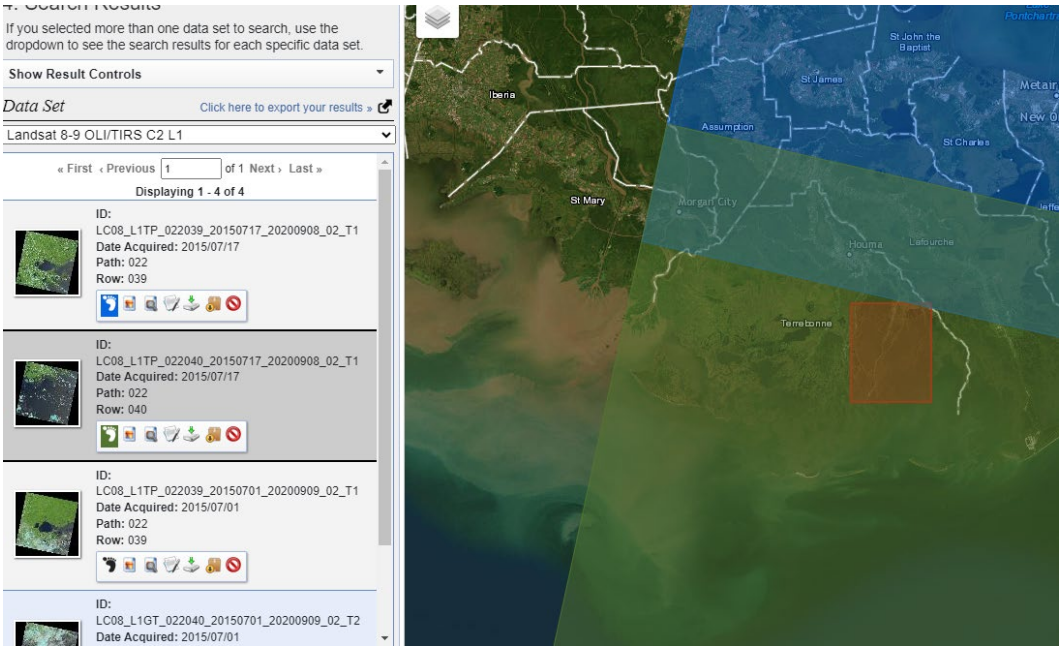
a) Go to https://earthexplorer.usgs.gov/ You will need to create an account and login.
b) Click the shapefile upload option with KMZ/KML. Click on “select file” and navigate to “watershed_grids/grid_kml” folder. Select a grid from the watershed that you think might have 3 overlapping tiles.
c) Add dates for the search criteria – 1 month will be sufficient (it does not really matter) I suggest choosing a month in 2015 to ensure it aligns with Landsat 7 and 8.



d) Under Data Sets Tab select Landsat Collection 2 Level-1 and click both L7 and L8-9. Then click on results.



e) Click the footprints on for all tiles and count number of overlapping tiles. In this case there are 2, so this is not a concern, but if there are 3 tiles then record the watershed, row, and column number. I believe the tiles are the same for landsat 7 and 8/9, but double check this with a few grids.



f) Upload the next sub-grid that is suspect and repeat until all sub-grids have been checked

Step 4: Upload training data as GEE Asset

a) Go to the training data folder (training_data\gee_repSites_allClasses) and upload each of the 5 training datasets associated with your watershed as a GEE Asset. Note for our test sites and model parameterization we will work with the 5 cross-validation data sets, but for the final training there will be 1 dataset for each watershed.
b) When the task is complete you can “refresh” under the asset tab and check that it is there. Click on it to make sure it imported properly

Step 5: Run GEE classification script

a) Open the “ccdc_classify_template_part2” script
g) Add your name, date, and comment at the top and hit “save as”. Save the script as “ccdc_classify_[watershed_number]_[cross_validation number]_[your_initials]
b) This script is set up to run the first cross validation dataset for Barataria but look at the parts that you will need to modify when you switch watersheds and cross-validation sets
c) Run the script and export 1 grid at a time, for each cross-validation dataset. After each run, you must submit the task for the file to export to your drive folder.

Step 6: Analyze results in MATLAB

a) Download all classification probability maps from your Google Drive folder to a place where you can access from MATLAB. Unzip all files and make sure there is a separate folder for each watershed and cross-validation dataset.
b) Open the “accuracy_assesment_template.m” matlab script. Add your name, date, and comment on the watershed. Save as and rename as “accuracy_assesment_[watershed_number]_[your initials].m”
c) This script is set up to run the first cross validation dataset for Barataria but look at the parts that you will need to modify when you switch watersheds and cross-validation sets
d) Run for each watershed and cross validation set. Save the confusion matrix as a “confMat_[watershednumber]_[cross-validation_number].mat” and take a screen shot of the figure produced. Past the figure into a word doc or powerpoint for later analysis.

Iterations

We are trying to determine the best parameters to run this model on our dataset, so we will perform several iterations of this process on our 5 representative sites. The iterations are outlined below. Note that some of these iterations require a change of segments (1st GEE script) and the classification (2nd GEE script) and some require a change in the classification only. You will likely run out of asset space with 1 set of segments, so it is wise to complete all classification iterations for a set of segments, exporting the results, and then deleting those segments prior to moving to the next segment iteration.

1) Run sample scripts on each watershed using default parameters in scripts (summarized below). **1 model run per watershed**
    a. Use all bands and indices in breakpoint and classification
    b. Use chi-squared = 0.97 on breakpoint detection
    c. Use 300 trees in random-forest classifier
2) IF accuracy between forest/scrub/emergent is low, test alternative training dataset (with scrub removed) using the same parameterization. **1 model run per watershed**
    a. I will provide the alternative training dataset and an additional MATLAB script to complete this.
3) Test alternative number of random forest trees. Only need to export segments once for each watershed. **2 additional model runs per watershed**. Begin extracting feature importance (I will provide GEE script and MATLAB code for this).
    a. Keep using all bands and indices in break point and classification
    b. Keep using chi-squared = 0.97 on breakpoint detection
    c. Test 150 and 500 trees (in addition to 300 that you already ran)
4) Test alternative chi-squared values in break-point analysis. Need to re-export segments each time. Note that you have already done 0.97, but will need to run the others with 3 different numbers of random trees. Therefore there are **6 additional model runs on each watershed**.
    a. Keep using all bands and indices in break point and classification
    b. Use chi-squared = 0.95, 0.97 (already completed), and 0.99 on breakpoint detection
    c. Test 150, 300, and 500 trees
5) Test different band/indices combinations. In addition to testing all bands and indices combined, you will repeat the runs using the bands only, and the indices only (plus the GREEN and SWIR1 in the breakpoint b/c of the Tmask algorithm). You should test all chi-squared and tree combinations, so there should be **18 additional runs per watershed**
    a. Use bands only in breakpoint and classification, and indices+GREEN&SWIR2 (for masking) in breakpoint and indices only in classification
        i. 5 Bands: BLUE, GREEN, RED, NIR, SWIR1, SWIR2
        ii. 6 indices: NDVI, MNDWI, NDWI, NPCRI, BSI, EVI, AEWInsh
    b. Test chi-squared = 0.95, 0.97, 0.99 in breakpoint analysis
    c. Test trees = 150, 300, and 500 in classification