Land Cover Classification Using Random Forest

**Overview:**

Random Forest is an ensemble learning algorithm for classification. It uses a large number of classification trees, each split created from randomly selected predictors and fully grown without pruning. Each tree casts a vote for the most popular class at a given input, and predictions are made by a majority vote. This project utilizes Random Forest for classifying land cover communities in the Lake Erie and Allegeny Partnership (LEAP) for Biodiversity.

**Inputs:**

1. Training Data:
   * PCAP community types
   * NatureServe (sp?)
   * Metroparks wetlands classifications
2. Predictors:
   * DEM derived data:
     1. SAGA Basic Terrain Analysis
     2. Terrain Position Index (200ft, 2000ft)
   * Historical Forest cover
   * Current Forest Cover
   * Soils
   * CIR Imagery
   * LiDAR
   * Other

**Methods/Workflow:**

1. DEM analyses: Running analyses on the DEM is the most time/resource intensive step
   1. Create a buffered grid to use as a cutline across the whole region. The buffer should be 2000 feet. The cut sections should not produce raster files larger than 1GB in filesize.
   2. Run DEMsplit10.bat with the DEM and the cutline grid as the input. The output will be *n* raster files, where *n* is the number of section in the buffered grid.
   3. Run runTerrain.bat in the directory where the cut DEM files are. This will produce 16 rasters of different types of analyses.
      1. If runTerrain fails on a particular file, it may be too big. You can use a similar workflow as this one to quarter it and run the terrain analysis individually, then merge those files back together.
   4. Run DEMsplit30.bat with the DEM and cutline grid as inputs.
   5. Run tpi200.bat where the 10 foot resolution DEMs are. (You might be able to run it on the whole thing, I forget)
   6. Run tpi2k.bat where the 30 foot resolution cut DEMs are.
   7. You now should have 16 files: tpi2k.tif, tpi200.tif, and 14 terrain analyses tifs.
2. Other predictors:
   1. Depending on the type of data, specific manipulations might be necessary.
3. Extent and Resolution:
   1. All the predictors must have the same extent and resolution. With all the files in the same directory, run extentresolution.bat
4. Removing residential and water bodies:
   1. Create a masking raster of the unwanted land types from NLCD.tif
   2. *More coming soon.*
5. Training Data:
   1. The training data needs to be prepared in a shape file of points. The training data comes from various sources, so you need merge all the data into one set of points. This set of points only needs one column with the community type label. Make sure that there is only one label for each community type (i.e. Beech Maple and BM need to be merged into the same label, BM). This can be done a variety of ways using GIS software such as Quantum GIS.
6. Random Forest:
   1. Method 1: R
      1. Open R and run it. Pros: produces more graphs and stuff. Can tune the data.
   2. Method 2: SAGA
      1. Create a shape buffer of training points.
      2. Import all the predictors and the shape buffer of the training points
      3. Run Geoprocessing > Imagery > Classification > Random Forest (ViGrA)
7. You now have a raster that is the predicted classifications for the region!