ISyE 6416 – Computational Statistics – Spring 2016
Final Project: "Big" Data Analytics
Proposal

Team Member Names:
Courtney Di Vittorio & Robert Woessner

Project Title:

Multi-temporal Classification of a Seasonally Flooded Wetland Using Optical Remote Sensing (Satellite) Indices.

Problem Statement:

The Sudd Wetland in South Sudan is a vital component of the Nile River Basin. During the dry season (Dec – Mar), the permanent wetland area is estimated to be 15,000 km^2. During the height of the rainy season in September, releases from Lake Victoria upstream augment the flows through the White Nile. As the high flows enter the Sudd, the water spills into the floodplains and expands outward. As a result of this combination of heavy rainfall and high flows, the flooded area of the wetland is believed to double in size, to approximately 30,000 km^2. Due to the high evapotranspiration rates in this semi-arid environment, only half of the water that enters the Sudd exits the wetland at the downstream end.

The desert countries downstream of the Sudd (Sudan and Egypt) would like to channelize the wetland and convey water downstream before it overflows and evaporates. However, there are nomadic communities who live within the Sudd and rely on the seasonal flooding cycle to regenerate the grasslands which feed their cattle. In addition, flow alterations to the Sudd would disrupt the sensitive ecosystem and the services it provides. Therefore, prior to diverting water from the wetland and conveying it downstream, decision makers should consider the detrimental impacts such alterations could have on the local people and environment.

In order to understand and quantify these impacts, a better understanding of the Sudd hydrology is needed. A key input required for the development of a hydrologic model of the Sudd is flooded area, which changes with time. In the past, areal extents of flooding were roughly estimated from only a few aerial images; ground data is not available given the scale of the wetland and the country's state of political instability. However, opportunities now exist to more accurately determine this parameter using remote sensing data from earth observing satellites. We will apply classification techniques introduced in the course to this satellite imagery in an attempt to distinguish between open water, permanently flooded vegetation, seasonally flooded vegetation, and upland vegetation.

Data Source:

The primary data source will be multi-temporal imagery from NASA's Terra Satellite which carries the Moderate Resolution Imaging Spectroradiometer (MODIS). From this sensor, the 500m spatial resolution 8 day composite land surface reflectance product will be used (MOD09A1). Each image from this product contains reflectance values from 7 bands (wavelengths), which have been atmospherically corrected.
The satellite re-visits the Sudd on a near daily basis, but due to frequent cloud cover a more complete image can be obtained by compiling the 'best' pixels over an 8 day period. However, even the 8 day composite product will have missing data from cloud cover. Each image contains a quality assurance layer which indicates whether clouds were detected or if there were any other quality issues. Any poor quality pixels will be eliminated from the analysis. The Terra satellite was launched in 2000, so images covering the Sudd from 2000-2014 will be analyzed. There are 46 images/year so over the 15 year period 690 images are available.

From the 7 spectral bands, various remote sensing indices can be derived. Because we are interested in mapping flooded area, we will use the Normalized Difference Vegetation Index (NDVI) which is a combination of the near infrared (NIR) and red bands and measures the amount of vegetation, and the Normalized Difference Wetness Index (NDWI) which is a combination of the green and short wave infrared (SWIR) bands and measures the amount of moisture. A major challenge in detecting flooded areas of wetlands from satellite imagery is that they are often covered in dense vegetation, and it can be difficult to distinguish between vegetation where the soil is wet from rainfall and vegetation where the soil is flooded. However, we believe we can distinguish between these two classes using a combination of the NDVI and NDWI and looking at the temporal trajectories of these indices versus single values in time.

In order to have some sort of ground truth data, we will use google earth imagery to identify geographic coordinates of locations with open water, permanently flooded vegetation, seasonally flooded vegetation, and upland vegetation.

Methodology:

Using the NDWI and NDVI values we will experiment with both clustering and supervised classification techniques. We can classify images individually for one instance in time or incorporate multiple images into the classification algorithm. For classifying images one at a time, we could use monthly mean values derived from the 8-day composite images.

In our analysis, we will use two types of learning methods: clustering and classification. They key difference between clustering and classification is that clustering is an unsupervised learning technique used to group similar instances on the basis of features whereas classification is a supervised learning technique used to assign predefined labels to instances on the basis of the training set of data.

Clustering and classification can seem similar because both techniques divide the data into subsets, but they are two different learning processes used for the purpose of getting information from raw data. The following is a description of the various techniques we plan to apply.

Unsupervised Clustering

1) *K-means*
   This is a relatively simple technique where we would not need to make any assumptions on the distribution of the data. The k-means algorithm would be applied to a single image at a time and the means of each class can be determined for the corresponding time.

2) *Expectation Maximization for Gaussian Mixture Models (EM for GMM)*
   Assuming the NDWI and NDVI values of each land class are normally distributed, we can apply this method to identify the mean and variance of each class for an image at one instance in time. In addition, the weights for the Gaussian mixture distribution can be determined which gives a soft classification for the pixels.

Once the pixels are assigned to a predetermined number of classes, we will need to manually determine which class each group represents. For both of these methods, we could include prior information on the proportion of the total area that is believed to belong to each class based on literature.

Supervised Classification

1) *Quadratic Discriminant Analysis (QDA):*
   Using the ground truth points, we will derive monthly mean distributions for each class. Subsequently each pixel within the images can be classified according to the proximity of the NDWI and NDVI indices to the mean values of the distributions.

2) *Logistic regression:*
   The statistical model will be the probability of being flooded, which is a soft classification. The coefficients of the model will be the NDWI and NDVI. Here we could include the indices from multiple time steps as the predictor variables.

The unsupervised clustering and supervised classification methods could also be combined. First the clusters can be determined using EM for GMM and the pixels with a high probability of belonging to a single class can be considered the ground truth points for that class. Subsequently these pixel groups can be used to derive the statistics in the QDA classification or can be used to fit the logistic regression model.

Principal Component Analysis (PCA)

Using PCA, we could reduce the number of images in our analysis. By vectorizing the images, stacking them in time or space, and deriving the principle components of the covariance matrices, we can reduce the full time series into a few images. These images should represent pixels that vary similarly in time or space with similar values. The images can then be classified using the previously mentioned techniques. The goal of classifying the images from the principal components is to utilize the full time series of data in one classification instead classifying multiple images from different instances in time. We could derive the principle components from the full time series, or split the data by month or by year.

Expected Results:

When classifying the images one at a time, the end result will be a time series of classified images. In order to determine the true class of each pixel, we might want to assign the pixel to the group that it is assigned to most frequently (for hard classifications) or take the maximum of the average classification weights (for soft classifications). For the multi-temporal classification results from the PCA or the logistic regression model, we can infer directly from the algorithm results.

Initially, we expect that the classification results using the principle components will more accurately classify the permanent land classes. However a time series of classified images might reveal more information. During the height of the rainy season, we predict that distinguishing between areas that are flooded from the river and areas that are wet from rainfall will be more difficult because everything will appear wet. However, by also looking at dry season imagery and the transition months (Nov/Dec and Mar/Apr) the extent of river flooding should be more apparent.