# The Use of Remote Sensing for the Prediction of Streamflow Statistics

**William Farmer, Anne Sexton, Brittney Veeck**

**CEE 194A: Remote Sensing**
**May 2, 2013**

## Estimating Streamflow in Drainage Basins

Streamflow is the volume of water in the stream flowing past a given point within a specific period of time (U.S. Environmental Protection Agency 2012). Because flow is directly related to the amount of water moving off the land into the stream, land use within the watershed is extremely significant. It is essential for developers and water resource managers to understand streamflow in any river basin. In the absence of streamflow records, those developers and water resource managers are left with the choice of simulating streamflow records or attempting to estimate streamflow statistics, such as the mean annual streamflow. In the latter case, regional regressions are typically used to estimate specific statistics. The purpose of this project is to explore the potential of remote sensing techniques for the prediction of streamflow statistics.

For this project we have decided to focus our attention on predicting the mean annual streamflow of 42 basins in the Southeast region of the United States. We chose the mean annual streamflow because it is both simple to calculate and widely used in real-world applications. It also stands as a proof-of-concept: If we can predict the mean with a high degree of accuracy, then there may be some potential in the prediction of higher-order statistics. We hypothesize that by classifying these basins and obtaining the areal coverage of different types of land-cover, we can predict certain streamflow statistics using regional regression techniques.

Regional regression is based on the theory that each region of basins can be classified by a unique regression equation. Vogel *et al.* (1999) showed that regional regression with drainage area, mean annual temperature and mean annual precipitation alone provided excellent predictive power. Farmer and Vogel (2013) later showed that more-advanced regressions could provide additional power, especially when variables related to land-cover and land-use were included. By classifying the land-cover in each of our basins we should be able to develop regional regressions that tie together the land-cover and the resulting streamflow behavior. This will both improve the ability to predict streamflow statistics and provide a framework for understanding how human behavior, as changes in land-cover, can affect streamflow.

We begin by presenting our analysis of the remote-sensing data available in our region. We then outline our classification techniques and discuss the performance of each classification method. We also discuss the struggles we encountered during our project and the lessons to be applied in future work. After, we present our regression analysis, quantifying the predictive power of land-cover in regional regression. Lastly, we conclude with a brief discussion of the potential for future research and the steps required.

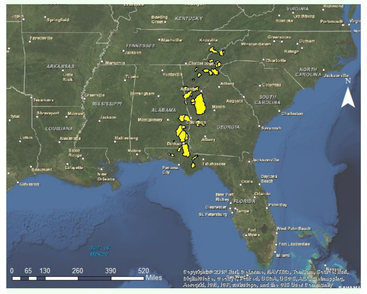


Figure 1: Extent of coverage of the 42 selected basins.

Figure 2: True Color Image of one of the 42 basins we classified.

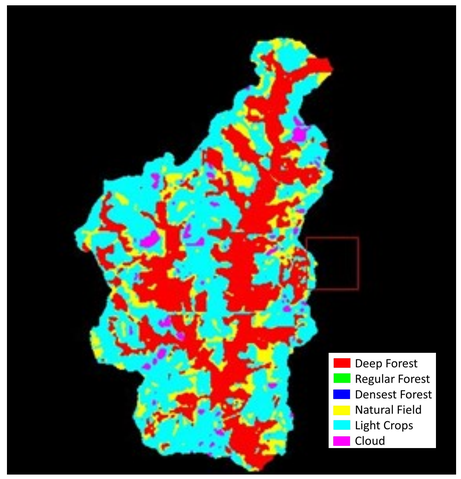


Figure 3: Kmeans Majority Image of one of the 42 basins.

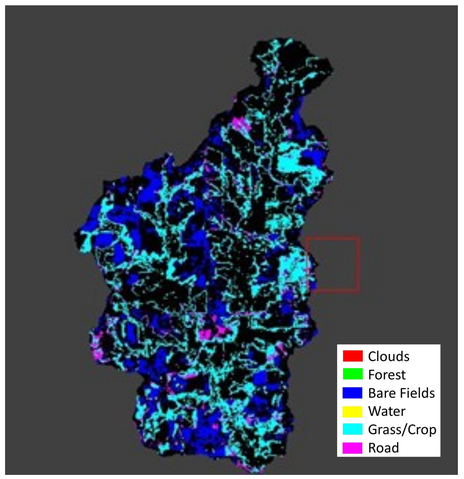


Figure 4: Parallelpiped Supervised Classification of one of the 42 basins.

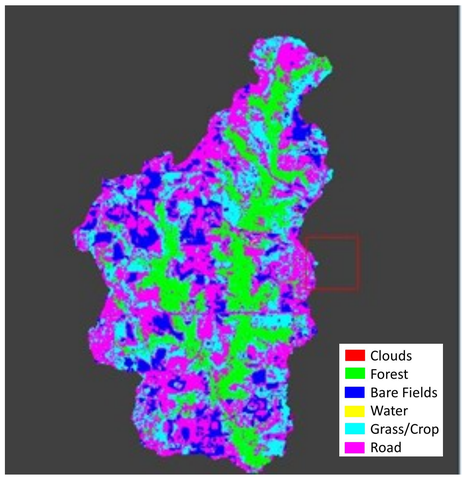


Figure 5: Maximum Likelihood Supervised Classification

## Remote Sensing Methods and Results

Once we decided on our basins, we needed to classify the land-cover. After importing the vector file of basins into ENVI, our first step was to create Regions Of Interest (ROIs) of all the basins in order to analyze the land-cover of each individual basin. These ROIs were then used to create a mask that would exclude any part of the Landsat scenes not covered by a basin in further analysis. This was done so as to reduce the processing time required to classify our regions.

The first classification method we used was an unsupervised K-means classification. We performed classifications several times to determine how many classes we thought were necessary and ultimately decided on six classes with five iterations. This approach seemed, to the eye, to provide the best balance between generalized land-cover classes and over-specified classes. Once we had the results of our classification, we used a majority analysis to smooth out the classification and try to create a more accurate image. The K-means classification is shown in figure 3 before and after the majority analysis. In the majority image, the red color represents deep forest, the green represents regular forest, the blue represents the densest forest, yellow represents natural field, cyan represents light crops, and magenta represents clouds.

Before we could judge the performance of the K-means classification, we needed to assess its accuracy. In order to perform any sort of accuracy assessment, we had to create ROIs that were representative of each of our classes. Because we needed to create these ROIs, we decided to also perform a supervised classification. We created two different sets of ROIs based on the same classes, forest, bare field, grass/crop field, water, roads, and cloud cover, to use for accuracy testing and supervised classification. The results of our K-means accuracy testing are shown in table 1.

We used one set of ROIs, as a calibration set, to perform a Parallelpiped Classification with a standard deviation of one. The result is shown below along with the true color image in figure 4. As you can see, the classification did not work very well. Because the standard deviation was so low, a large number of pixels were left unclassified. We found that if we increased the standard deviation, the image was classified without much specificity. This may indicate that the classes we were considering were too closely lumped to allow for sufficient discrimination.

Because our first supervised method did not seem to perform well, we decided to perform a Maximum Likelihood Classification with our calibration set of ROIs. With this classification methodology, ENVI could not differentiate between roads and bare fields. To remedy the confusion, we edited the original set of ROIs to try to distinguish the two classes better. We then ran the Maximum Likelihood Classification with the revised calibration set of ROIs. The result is shown in figure 5 along with the true color image. This revision resulted in a classification that, to our eyes, more accurately captured the main elements of each basin: forest, agricultural fields and human development.

After comparing the images by site, we decided to perform an accuracy assessment of all our classification methods. To conduct this assessment we developed an independent set of validation ROIs. We then produced a confusion matrix for the Maximum Likelihood Classification (table 3(a)) as well as for the Parallelpiped Method (table 3(b)). We also attempted to provide a confusion matrix for the K-means classification (table 3(c)), but we were unable to completely match our supervised ROIs with the computer-generated classes from the unsupervised technique. Instead we had to rely only on the confusion of three classes.

The results of the Kappa statistic and overall accuracy are summarized in the table 1**.** The closer the accuracy is to 100% and the closer the Kappa statistic is to 1, the better the classification. The accuracy is very low in both the K-means and Parallelpiped methods but for two different reasons. The ROIs we created to compute the confusion matrix did not match up with the classes that ENVI generated during the K-means classification. As a result, the accuracy is low because some of the classes were not compared when computing the accuracy matrix. The Parallelpiped classification failed to classify many pixels because of the low standard of deviation we assigned and, as a result, had a low accuracy.

Because we could not adequately assess the accuracy of the unsupervised techniques, we decided to run our regression using both the maximum likelihood classification and the K-means classification. The maximum likelihood classification had the highest accuracy, but we were unsure of how to determine the actual accuracy of the K-means method. In order to run the regression, we computed the percentage of each basin ROI that was covered by each class for each classification methods. These percentages, representing the areal coverage of land-cover in each basin, served as the explanatory variables considered below.

Table 1: Accuracy Test Results of Supervised and Unsupervised Classification Methods (The K-Means assessment is based on a subset of classes.)

| Classification Method | Overall Accuracy | Kappa Value |
|---|---|---|
| K-means | 46%** | .35** |
| Parallelpiped | 33% | .27 |
| Maximum Likelihood | 76% | .69 |

Sources:

Farmer, WH, and RM Vogel. 2013. Performance-weighted methods for estimating monthly streamflow at ungauged sites. *Journal of Hydrology* 477: 240-250.

Falcone, J. 2011. GAGES-II: geospatial attributes of gages for evaluating streamflow. Database. US Geological Survey, Reston, VA. http://water.usgs.gov/lookup/getspatial?gagesII_Sept2011

U.S. Environmental Protection Agency. *5.1 Stream Flow.* March 6, 2012. http://water.epa.gov/type/rsl/monitoring/vms51.cfm (accessed April 29, 2013).

Vogel, RM, I Wilson and C Daly. 1999. Regional regression models of annual streamflow for the United States. *Journal of Irrigation and Drainage Engineering* 125.3: 148-157.

U. S. Geological Survey. 1995. http://landsat.usgs.gov/ (accessed April, 2013).

## Regression Analysis

For this analysis we used weighted least-squares regression to predict mean annual streamflow on the basis of several explanatory variables. Each observation was weighted by the length of the record used to compute the mean annual streamflow. In addition to the percent areal coverage of each land-cover class, we used drainage area, mean annual precipitation, and mean annual temperature as explanatory variables. We then developed a computer algorithm to conduct an F-test on each combination of explanatory variables. An F-test is used to quantify the probability that each set of variables is jointly significant. By conducting an F-test on all combinations of variables, we can ensure that the final model is the most parsimonious, using only explanatory variables that provide significant predictive power.

The model developed is a mixed power-law model. Rather than using a traditional linear model, we considered the regression of the logarithm of mean streamflow against the logarithms of drainage area, precipitation, temperature, and the real-space percentages of areal coverage. The resulting model is of this form:

$$\ln(\mu) = \beta_0 + \beta_1 \ln(A) + \beta_2 \ln(P) + \beta_3 \ln(T) + \sum_{i=1}^{N} \alpha_i C_i$$

where $\mu$ is the mean annual streamflow, $A$ is the drainage area, $P$ is the mean annual precipitation, $T$ is the mean annual temperature, $N$ is the number of classes considered, $C$ is the areal coverage of the $i$-th class, and the $\beta$s and $\alpha$s are regression coefficients. Models were assessed on the basis of adjusted-$R^2$ (values closer to one are better) and Nash-Sutcliffe efficiencies. The Nash-Sutcliffe efficiency is the best measure of how well a model is performing by converting the adjusted-$R^2$ into a real-space interpretation of predictive capability, with ideal values close to one.

We conducted our regression analysis on the supervised maximum-likelihood method and the unsupervised K-means approach. Because we could not reconcile the classes of these two methods, we hypothesized that it would be worthwhile to consider both regressions, as the unsupervised techniques have found something that was not captured by our supervised techniques. In order to avoid the confounding influence of clouds, we excluded basins that were classified with more than 2.5% of the basin being covered by cloud. Filtering cloud coverage resulted in 27 qualified basins. In the end neither classification scheme provided groundbreaking results, but the unsupervised method does show promise for future research.

When we considered all explanatory variables for a regression of the supervised classification, we found a high adjusted-$R^2$ (0.9921), but this is a bit misleading. Because the areal coverage across classification schemes must always sum to 100%, including them all in the regression results in an inflation of variances and biased, inconsistent estimates of the coefficients. This is evidenced by the non-normality of the residuals, with a normal probability-plot-correlation-coefficient (PPCC) of 0.9401. (In the terms of a hypothesis test for normality, normal residuals would have a PPCC below 0.9619 only 5% of the time.) We decided we could not trust this first regression because the residuals are not normal. By conducting a series of F-tests, we found that the simplest model that could not be rejected (p=0.10) was one that contained only the climate information. The resulting adjusted- $R^2$ was 0.9894 and the Nash-Sutcliffe efficiency was 0.9773. This suggests that there is little added value to including our supervised classification.

Interestingly, the results were not the same for the unsupervised classification. The general model, with all explanatory variables, was also flawed by the variance inflation, noted above. As such, we were again forced to conduct our enumerative series of F-tests. The most parsimonious model included all the climate variables and two classes of the unsupervised classification scheme. Upon closer examination, these classes seemed to match up to agricultural fields (Class 5) and a subset of the forests (Class 2). This forest subset was characterized by dense, interior stands of foliage. The resulting adjusted- $R^2$ was 0.9951 and the Nash-Sutcliffe efficiency was 0.9915. This shows that the classification provided by the unsupervised techniques captured a unique signature that is essential to the prediction of mean annual streamflow. The full set of coefficient estimates are shown in the following table.

Table 2: Regression coefficient estimates and statistics for the prediction of the natural logarithm of mean annual streamflow.

| | (Intercept) | ln(A) | ln(P) | ln(T) | Class 2 | Class 5 |
|---|---|---|---|---|---|---|
| Estimate | 93.3013 | 1.0042 | 2.5726 | -18.8792 | 0.0048 | 0.0123 |
| t-Statis- | 4.5888 | 61.1263 | 15.1882 | -5.1970 | 2.3689 | 3.7409 |
| p-value | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0275 | 0.0012 |