Empirical Modeling of Remotely Sensed Data at Regional to Continental Scales

Richard D. Robertson
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
rdrobert@uiuc.edu

Praveen Kumar
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
kumar1@uiuc.edu

Peter Bajcsy
National Center for
Supercomputing Applications
University of Illinois at Urbana-Champaign
pbajcsy@ncsa.uiuc.edu

David K. Tcheng
National Center for
Supercomputing Applications
University of Illinois at Urbana-Champaign
dtcheng@ncsa.uiuc.edu

Abstract

A continental scale dataset was assembled to examine the drivers of greenness indices. Easily parallelized algorithms for Ordinary Least Squares linear regression and a binary regression tree were implemented and used for the analysis. The most important drivers were found to be long and shortwave radiation, precipitation, elevation, and soil pH. This analysis shows that it is possible to perform empirical modeling with large datasets and thus the archives of remotely sensed data can and should be analyzed to shed light on models of large scale natural processes.

1. Introduction

Ever since the advent of earth observing satellites, an extremely large amount of data has been available for scientific purposes. However, the sheer volume of the data has in practice tended to limit its use to descriptive methods. Now, there are sufficient computational resources ranging from personal computers to supercomputers able to tackle these large datasets.

This analysis demonstrates a way of considering the prediction of greenness indices at the continental scale. Our dataset was constructed for the continental United States at 1km resolution providing about 6 million pixels for analysis and occupying about 1.3GB of storage. This is about the practical limit that can be analyzed on a single computer in an "on demand" way, *i.e.*, the results are available in a matter of minutes or hours without resorting to high performance computation resources that are typically shared with other users which increases the wait-time before the job is executed. While the results reported here were compiled using a laptop, the algorithms were constructed as single machine simplifications of parallelized code.

2. Methods

We employ two standard approaches for predictive empirical modeling of a single continuous output based on a collection of explanatory variables.

2.1. Linear ordinary least squares

The first is Ordinary Least Squares (OLS) linear regression. In this case, we use the standard classical assumptions of homoskedasticity and normality for the errors. The systematic portion of the model is taken to be linear. That is, letting Y be the dependent variable to be modeled/predicted, X be the explanatory variables with an initial column of ones, β be the collection of coefficients and a constant, and ε the error terms:

$$Y = X\beta + \varepsilon \tag{1}$$

Under the classical assumptions (which may not, in fact, hold), the OLS estimator is the Maximum Likelihood Estimator and is computed by

$$\hat{\beta}_{\mathrm{OLS}} = \left(X'X\right)^{-1} X'Y \tag{2}$$

Obviously, using a linear model to approximate the phenomenon under consideration is quite simple and can be improved upon. We include it because it is a standard approach that is useful for comparison.

2.2. Regression trees

The second technique is a kind of binary regression tree. The basic idea is to devise a set of nested conditional statements, each based on a single variable, that subdivide the characteristic space. Each resulting region is assigned a value corresponding to the prediction of the dependent variable. The combination of the rules and values defines a piecewise constant function that approximates the phenomenon being modeled.

The rules can be thought of as a tree diagram starting with a single decision based on a single variable and a threshold as is shown in **Figure 1**. If the variable is less than the threshold, we consider the "left branch," otherwise we proceed down the right branch. The simplest possible tree would be to stop after the first decision resulting in one of two possible predictions. Here we will refer to such a tree as having two levels: the first is the decision while the second has the predictions in its "leaves."

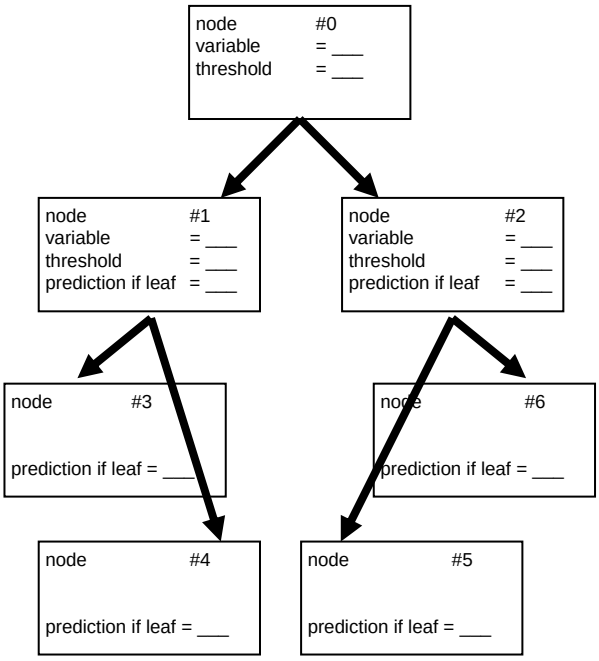

Figure 1. Diagram for a three level tree. Within a decision, if the variable is less than the threshold, the left branch is followed, otherwise, the right branch.

The predicted values are assigned by applying the rule (assume for the moment we have already determined a good rule) to all of the data we are using to estimate the model and finding which examples end up in each leaf. In our particular specification, we compute the sample mean of the dependent variable over the examples that end up in each leaf and record the value as the prediction for that leaf.

The rules are chosen by searching among a restricted set of alternatives. Here, all of the rules have the form $X_k < \mu_k$ where X_k is the value of the kth explanatory variable and μ_k is the associated threshold. We used the sample means of the examples being fed into a particular decision as the candidate thresholds for that decision. The rule chosen is the one which results in the lowest total error for the examples fed into the decision. We used the summed squared error function:

$$\text{Error} = \sum_{\text{all pixels}} \left(Y_{\text{actual}} - Y_{\text{predicted}}\right)^2 \qquad (3)$$

Clearly, different flavors of the same approach could be obtained by using different thresholds and predictions (*e.g.*, the median) and error functions (*e.g.*, absolute deviations).

The trees can be expanded to multiple levels by splitting each branch of a level into two new branches leading to leaves on a new level. The decision rules and predictions are similarly assigned based on the examples that filter down each individual path. Notice that the total number of paths and hence predictions grows exponentially with the number of levels employed. For example, using only the first two levels results in two unique predictions based on one decision while using three levels provides four predictions based on three decision rules (one in the first level and two in the second). In general, using n levels will result in a total of $2^{(n-1)}$ uniquely predicted values based on $2^n - 1$ decision rules.

3. Data

We assembled 29 explanatory variables and the Enhanced Vegetation Index (EVI) on a common projection and resampled them to a common 1km grid.

The vegetation index data is taken from MODIS Terra EVI (MOD13Q1 version 4). These datasets are available at 250m spatial resolution and a temporal scale of 16 day maximum value composite (MVC) in order to reduce the effect of clouds and aerosols. All the water and bad quality pixels from the datasets were removed using the MODIS EVI QA/QC (quality assurance and quality control) mask available with the EVI datasets. This MVC EVI data was averaged on pixel by pixel basis for the month of April and a minimum threshold of 0.1 was applied to remove the pixels with very low vegetation or no vegetation.

Several topographic variables from the HYDRO1k suite of datasets based on the GTOPO30 DEM were included: elevation, slope, northern and eastern components of aspect, and the compound topographic index (CTI). Elevation data is in meters. Slope is expressed in degrees time one hundred varying from 0 to 9000. Aspect varies from 0 to 360 and is

the direction in which the slope is facing. This was converted to northern and eastern components by taking sine and cosine. Additionally, we included a variable reflecting the distance to the nearest stream.

Various soil properties were included and were taken from the State Soil Geographic (STATSGO) database, which is at a 1km spatial scale. The variables were pH, available water capacity (cm), bed rock depth (cm), bulk density (g/cm^3), permeability (cm/hr), and percentage sand, silt and clay. Most of the soil properties are averaged over the 11 vertical soil layers to obtain a single value per pixel.

Meteorological variables such as precipitation (mm), day and nighttime surface temperatures (°C) and long and shortwave radiation (W/m^2) were obtained from National Land Data Assimilation (NLDAS) project. Data is available on an hourly basis that is averaged over a month with a spatial scale of 12 km. The datasets from NLDAS are model outputs of the Land Data Assimilation System (LDAS) with an operational numerical weather prediction (NWP) system. The model integrates past forecasts with observations from various data sources to improve the performance, *e.g.*, it uses radar measurement as well as rain gauge observations to improve the model output for precipitation.

The 1 Km Global Land Cover Product from the University of Maryland provided land use/land cover variables. Of the 15 designations in the dataset, we kept pixels for 11 and constructed indicator variables for 10 to avoid identification problems in the regression model. The retained categories were Evergreen Needleleaf Forests, Evergreen Broadleaf Forests, Deciduous Broadleaf Forests, Mixed Forests, Woodlands, Wooded Grasslands/Shrublands, Closed Bushlands or Shrublands, Open Shrubland, Grasslands, Croplands, and Barren (omitted indicator variable).

After dropping missing values and inappropriate land use categories (*e.g.*, water, urban areas), there were a total of 6,252,945 examples composed of 29 explanatory variables and the EVI.

4. Results

The first step in the analysis was to divide the data into three parts. The first pseudo-randomly selected part was 48% of the original data to be used for estimating the models. This is referred to as the "Training Set." Another 32% was set aside as a "Validation Set" to be used in determining a good specification and to assess overfitting. The final 20% of the data was set aside as a "Testing Set" to provide an unbiased assessment of the final models performances since the Validation Set is used to choose the specifications.

4.1. All variables

We began by using all the variables for both the linear model and the regression tree. We trained the tree to 20 levels. Using the estimated models, we computed the summed squared error for all three subsets of our data. For comparison, we transformed this into the root mean squared error. This quantity represents the typical error standard deviation around the predicted values and provides a summary measure of the models performance. These values are shown in **Figure 2** with the OLS results plotted as horizontal lines for comparison with the tree results.

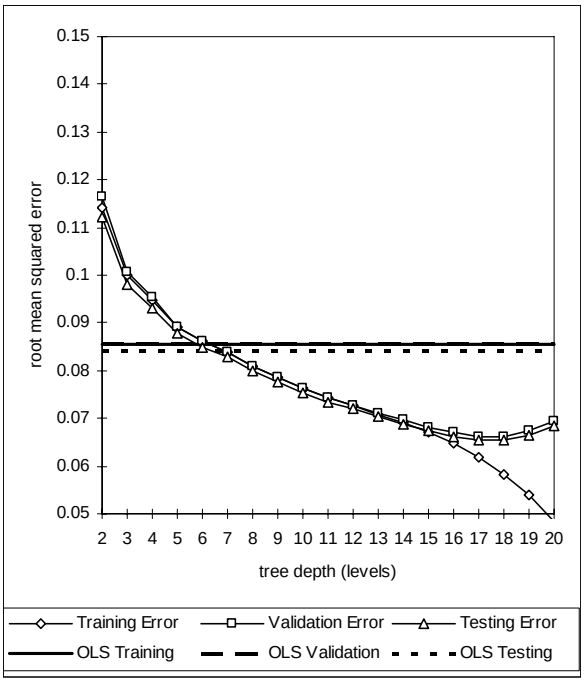

Figure 2. Summary performance, models estimated using all variables.

The summary statistics reveal a few interesting things. The OLS training and validation results are virtually identical and are roughly equivalent to a tree of depth 6 (based on 5 decisions for any particular example). The tree performs better than the linear model and begins to overfit at about depth 14 with the validation error bottoming out at about 17 levels. The OLS error is about 0.085 while the trees uncertainty is about 0.065 . The EVIs in the dataset range from about 0.1 to 1.0 with a mean of about 0.289 and standard deviation of about 0.146.

We next examined each model for evidence concerning which variables were most important.

For the OLS model, two criteria were considered. First was the magnitude of the effect on the predicted EVI of a

change in the explanatory variable. Since the model is linear, this is accomplished by inspection of the coefficients. We rescaled each coefficient by dividing by the sample standard deviation of the corresponding variables. This rescaled coefficient represents the expected change in the predicted EVI when the explanatory variable changes by one standard deviation of the observed distribution of values for that variable.

However, it is not enough to look for large values for the rescaled coefficients because the data may not support the estimate very well. Due to the statistical assumptions behind the model, we can compute the *t*-statistic for each coefficient in order to test the null hypothesis that it is equal to zero. Both the standardized coefficients and the *t*-statistics are plotted in **Figure 3**. Of course, due to the large amount of data, almost all of the parameter estimates were "statistically significant." However, some were more so than others. One additional thing to note is that the land use/land cover was included as a series of indicator variables. The indicator for barren was excluded, and hence the estimated coefficients show the effect of the stated land use as compared to the omitted category: barren. Hence, the land use coefficients are, for the most part, quite different from zero (indicating they are different from "barren"), but quite similar to each other (indicating they have similar effects). Considering this combination of magnitude and testing for magnitude illuminates the following variables as most important: longwave radiation, shortwave radiation, precipitation, day and nighttime temperatures, elevation, and soil pH. Closer inspection indicates that the day and nighttime temperature coefficients look similar but with opposite signs. Since it is reasonable to expect that the two temperatures are related, this indicates that their effects seem to cancel each other out and thus only one (or neither) of the variables is actually necessary.

The tree was examined using the method of White [1]. Her approach was to construct an index based on how often a variable was used for a decision and rewarded occurrences higher in the tree more than lower decisions. The resulting values for this tree are shown in **Figure 4**. Employing this notion of Amandian Relative Global Dominance, we found that longwave radiation was the most influential. The other important variables were: precipitation, shortwave radiation, nighttime temperature, elevation, and soil pH.

As a test to see if the identified variables were, in fact, the most important, we split each of the datasets into two pieces keeping all the observations in the same order. The "Good" data contained all the variables identified as important, specifically: longwave radiation, shortwave radiation, precipitation, nighttime surface temperature, elevation, and soil pH. The "Bad" data contained the remaining explanatory variables.

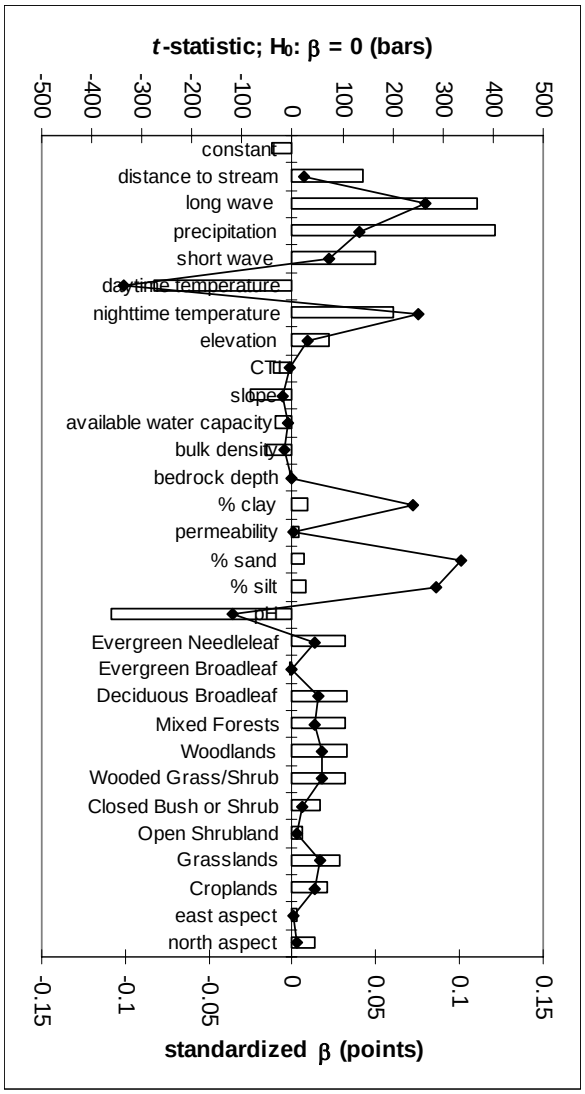

Figure 3. Standardized coefficients and t-statistics for linear OLS model (all variables).

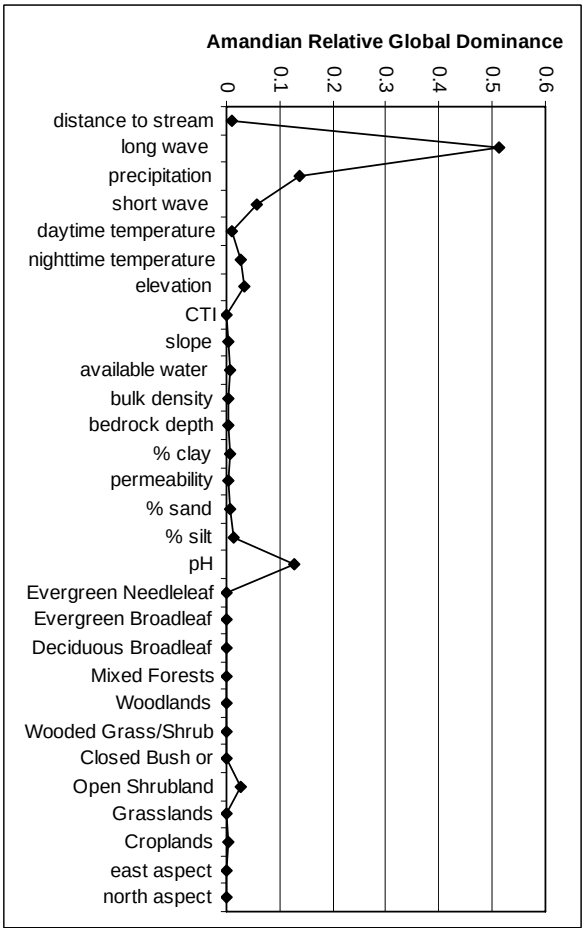

Figure 4. Amandian relative global dominance for 17 level regression tree (all variables).

4.2. Good and bad variable results

Again, we estimated a linear model and trees to depth 20 for both sets of explanatory variables. The resulting summary performance statistics are shown in **Figure 5**.

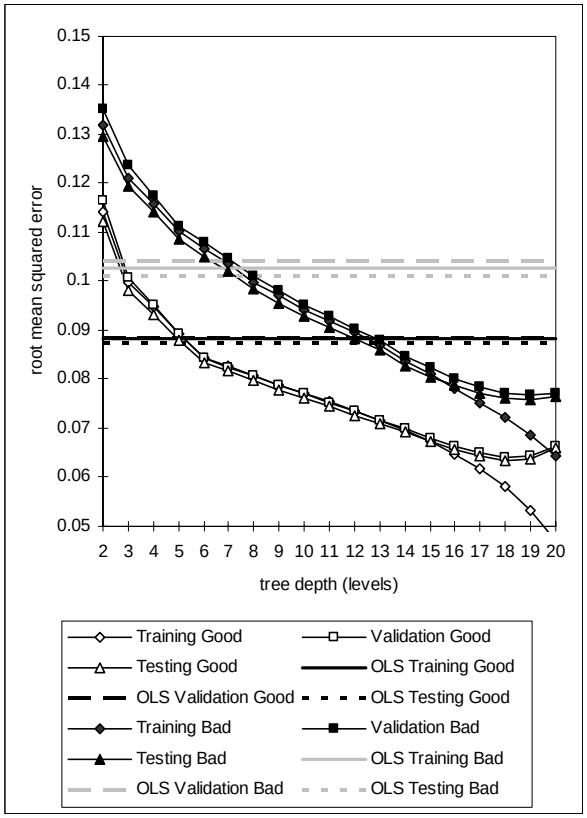

Figure 5. Summary performance, models estimated using "Good" and "Bad" variables.

The OLS results behave as expected. The restricted "Good" model performs slightly more poorly than the unrestricted "All" model. The "Bad" model results in a much worse fit (about 0.102 for the Bad as opposed to the 0.085 and 0.088 for the All and Good). The testing error is also much higher than the models based on the "All" and "Good" variables. Taken together, these results indicate that the "Good" variables really are important from the perspective of the linear model and account for almost all of the models predictive power.

The story is similar with the regression trees. The "All" tree and the "Good" tree have very similar errors (within 0.003 of each other). For some levels (6, 7, 17-20), the Good tree actually performs better than the All tree. Apparently, there are a few bad decisions made when all the variables are available which are avoided with the restricted model. The "Bad" tree performs much worse indicating that

those variables are not merely variations on the Good variables.

5. Conclusions

The availability of cheap computational resources enables us to perform empirical modeling exercises on the vast wealth of remotely sensed data that has been piling up for decades. In this analysis we considered the factors driving greenness indices using a continental scale dataset. We were able to successfully perform the analysis on a laptop using algorithms that easily parallelize.

We found that the most important drivers are long-and shortwave radiation, precipitation, nighttime surface temperature, elevation, and soil pH.

Future work should extend this analysis in several directions. First, finer grained data should be employed: the EVI is available at 250m resolution and should be used. Second, additional variables should be included to reflect human economic control of parts of the biosphere to gauge their relative importance. Examples of such variables are population densities, cost of access indices, protected area designations, and site specific prices. Third, more varieties of models should be employed such as neural networks, composite "forests" of trees, or polynomially interacted regression models.

Computational power is now cheap and algorithms capable of handling large datasets are available so we should not fail to sift through the mounds of existing data for evidence to support or undermine our models of large scale natural processes and their interactions with human activity.

6. Acknowledgement

Many thanks to Vikas Mehra for his instrumental work on data preparation. This work was made possible by a grant from NASA entitled "Data Mining for Understanding the Dynamic Evolution of Land-Surface Variables: Technology Demonstration using the D2K Platform."

References

[1] A. B. White and P. Kumar. Dominant influences on vegetation greeness Part I: the Blue Ridge ecoregion. *J. Geophys. Res. Biogeosciences*, 2006. in press.