Data Driven Discovery from Satellite Remote Sensing: System Development and Analysis of Vegetation Indices

Praveen Kumar
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Richard D. Robertson
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Amanda White
Environmental Geology and Spatial Analysis
Los Alamos National Laboratory
Los Alamos, NM 87545

Vikas Mehra
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Peter Bajcsy
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Pratyush Sinha
Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Wei-Wen Feng
Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801

David K. Tcheng
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, IL 61801

***Abstract*-This paper describes the GeoLearn system for preparing remote sensing datasets. Three applications are given for analyses using collections of remote sensing data for modeling. The Blue Ridge ecoregion is used as the first area for understanding the influences on greenness indices. The scale of analysis is greatly increased by considering a similar approach using data from the entire continental United States. Finally, clustering algorithms are applied to various land surface variables to look for simple relationships.**

I. Introduction

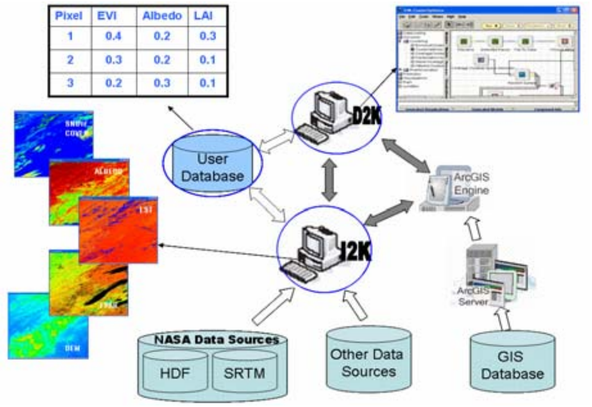
This research revolves around processing and analyzing remotely sensed land surface variables. We briefly review the capabilities of the data processing tools that we have developed and then discuss several investigations that apply data-driven modeling techniques to discover relationships between topography, climate, soil properties, and vegetation indices.

Understanding these relationships requires assembling data at regional and continental scales. This in turn requires the ability to handle very large volumes of data that are dynamically evolving such as weather, emissivity, vegetation, as well as static features of the landscape. Models based on investigations of these data will improve our understanding of how the variables evolve and influence one another. The benefits may also extend into other areas since the models could be employed to refine the approximations in landscape and climate simulation models.

II. The GeoLearn Wizard

One of the primary challenges in this type of research is dealing with disparate kinds of data. There is a large volume of georeferenced data, but each dataset is potentially recorded in its own projection and resolution (both temporal and spatial) with or without information about data quality, and in a variety of formats. We are developing the GeoLearn wizard as a software tool to streamline the process of dealing with the challenges of bringing these data together with a special emphasis on handling the remote sensing data sets collected by the MODIS instruments on the TERRA and AQUA platforms.

The GeoLearn wizard forms a graphical interface that employs tools from the Im2Learn image and GIS package developed by the ISDA group at NCSA



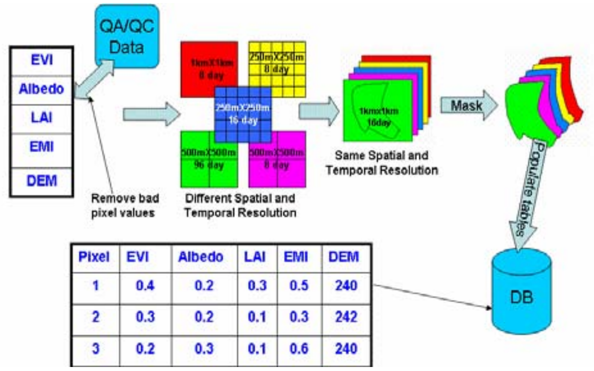
Pixel	EVI	Albedo	LAI
1	0.4	0.2	0.3
2	0.3	0.2	0.1
3	0.2	0.3	0.1

Fig. 1. Diagram showing the general architecture of GeoLearn.

(http://isda.ncsa.uiuc.edu/Im2Learn/doc/) along with other tools in ESRI's ArcGIS Engine (http://www.esri.com/software/arcgis/arcgisengine/). The general architecture is shown in Fig. 1 and the approach to processing the data is shown in Fig. 2.

The first major function of GeoLearn is to ingest the raw raster data and perform several preprocessing steps. These include the obvious needs for spatial and temporal resolution adjustments to match up datasets on different grids and time intervals; reprojecting data onto a common map projection; and mosaicking multiple tiles into a single maps. One of the more novel capabilities is the ability to use the Quality Assurance/Quality Control information provided with the MODIS products to mask off unwanted data such as water, ice, and snow or cloud and shadow influenced measurements. Similarly, GeoLearn allows vector data designating polygons to be loaded, reprojected, and layed over the raster data. These polygons can be selected and used to create another mask for selecting only the pixels within their boundaries. This allows the user to restrict the final dataset to a limited geographic area, *e.g.*, particular states or ecoregions.

The second major function of GeoLearn is to convert the processed data (now on a common grid/projection/time-scale with undesired data masked off) into data structures ready for empirical modelling. This can take a couple of forms. The most easily employed is the D2K Table which is a tabular data structure that is ready for use within NCSA's Data2Knowledge (D2K) datamining environment (http://alg.ncsa.uiuc.edu/do/tools/d2k/). D2K has a wide variety of knowledge discovery algorithms already prepared and coded. GeoLearn is set up to be able to easily access these algorithms for exploring and modeling the constructed dataset. The data can also be exported as binary files for later use.



Pixel	EVI	Albedo	LAI	EMI	DEM
1	0.4	0.2	0.3	0.5	240
2	0.3	0.2	0.1	0.3	242
3	0.2	0.3	0.1	0.6	240

Fig. 2. Workflow showing the approach to processing the data.

Finally, for a few algorithms, GeoLearn provides seamless visualization of the results of predictive models. For example, a regression tree algorithm is implemented and GeoLearn can visualize maps of predicted values and errors for different levels of the tree. Additionally, it provides some graphical interpretation aids such as showing which variables are most important for the model at each location.

III. Initial Application: Predictive Model of Blue Ridge (Level III) Ecosystem

The Blue Ridge Ecosystem provided an initial region in which to investigate the dominant influences on vegetation indices. For a more detailed description of this portion of the research, see [1]. The Blue Ridge is a primarily forested mountainous area in the Appalachian Mountains of the eastern United States. The ecosystem is one of the most biologically diverse temperate broadleaf forests in the world, covering an area of about 46,600 square kilometers with a variety of geologic features stretching from Georgia to Pennsylvania.

The vegetation index analyzed was the Enhanced Vegetation Index (EVI) from the MODIS Terra for April to September, 2000 to 2004. For each 250 meter pixel in the Blue Ridge region, a single EVI value was computed for each month by taking the mean of all the 16-day maximum composite values within the month for all five years. This provides six maps of average EVI disaggregated by month.

Several types of explanatory variables were assembled. By using the explanatory variables to predict the EVI values and examining the model and its predictions, we can learn which variables are more and less important in influencing the vegetation index.

The first type of variables were land cover indicator variables which encoded the presence (value = 1) or absence (value = 0) of categories from the USGS National Land Cover Dataset. Some categories were excluded from the analysis: water, urban areas, bare rocks, *etc*. Three of the ten categories accounted for the vast majority of the land cover: deciduous forest and evergreen forest along with some pasture/hay.

Second, several topographic attributes were included. These included elevation and its derivatives slope and aspect (the direction the slope is facing). In addition, two variables reflecting water availability were constructed: the compound topographic index (CTI) based on the slope and upstream contributing area, and the distance to the nearest stream.

A collection of soil properties comprised the third group of variables. Many of these were thickness-averaged over all soil layers, when applicable. The particular properties were: percent sand, silt and clay, permeability, total bulk density, pH, percent available water capacity, and depth to bedrock.

The final group of variables were meteorological: unfrozen precipitation, incoming short and longwave radiation flux, and day- and night-time temperatures. Each variable was averaged over the days in each month to produce a single value for each month.

The data were analyzed by training binary regression trees to predict the EVI based on the explanatory variables which were resampled to have the same resolution as the EVI map. A regression tree was trained for each month.

A regression tree is a type of model for continuous functions based on the idea of building a parsimonious piece-wise constant approximation. The edges of the pieces are defined by thresholds associated with particular explanatory variables that determine which value of the predicted variable is to be assigned. The entire model is known as a “tree” because the threshold decisions are built up by considering a series of conditional statements which each define a split into two possible alternatives. Each of those alternatives has its own threshold-and-variable conditional statement which splits into two more possible alternatives. When depicted as a diagram, the model appears as a tree with several levels beginning with a single decision at the top with ever increasing numbers of branches reaching downward. By increasing the number of layers used in the tree, the researcher can build a more detailed approximation. The process of determining which variables and thresholds should be associated with each decision is known as “training” the tree.

The flexibility in decision trees means that care must be taken to avoid obtaining unrealistically optimistic results. This is typically done by splitting the data into two or three distinct sets before beginning the analysis. The “training” set is usually the largest and is used for training the models being used (here, decision trees of various depths). The second set is called the “validation” set and is used once the models have been trained. The models are used to make predictions for the validation data and compared to the actual values. This allows an assessment how the models perform on data *not used* in their creation. Based on these error assessments, different models can be compared to determine the best one. With decision trees, the most important comparison is to determine the number of levels that performs best because an excessively deep tree will perform poorly on the validation set. Finally, sometimes a third “testing” set is kept separately from the training and validation sets in order to obtain a final unbiased estimate of the error or performance of the model that is chosen as the best. A third independent set of data is required because the validation set has already been used to compare between candidate models and thus may result in an optimistically biased error estimate.

Examination of a trained tree provides insight into which explanatory variables influence the dependent variable the most. Clearly, if a particular variable is never used to make a “decision,” it is probably not a very important influence. On the other hand, variables that are important will appear frequently in the tree. Furthermore, we might assume that decisions made higher in the tree (the first few decisions on which the rest depend) are in some sense more important than the last few. These ideas can be used to define dominance scores that reflect the overall importance of a variable in the tree as compared to the others based on how often it appears and weighted by which levels it appears in. Similarly, the values for a single pixel can be run through the tree in order to see which variables are used in the decisions resulting in the final prediction. These provide both general assessments of the phenomenon as well as the local behavior for particular situations.

The relative influence of each type of variable on the vegetation index was determined by computing the dominance scores for each and pooling them within the land cover, topography, soils, and meteorological categories. These summary Amandian Relative Global Dominance measurements are plotted in Fig. 3. Overall, the soil properties do not appear to be very important and do not change in their importance as the growing season progresses. Meteorology seems slightly more important and its importance only changes in that it appears to be less important during June and slowly increases in importance. Land cover appears to be the dominant driver during April (roughly as important as the other three types of explanatory variables combined), is conspicuously unimportant in May, and assumes a moderate importance that gradually declines during the remainder of the summer. The land cover is probably most important early in the season because of the differing growth patterns of evergreen trees as compared to deciduous trees in early spring. Finally, topography begins April as secondarily important to land cover, but switches in May to become the overwhelmingly dominant set of variables and continues to be so throughout the growing season.

Another way to consider the relative importance of different drivers is to examine which variables are used in the decision tree for each location on the map. In Fig. 4, the maps show which types of variables are most important at each location by month. Based on the general measures of importance, it is not surprising that in May, the majority of the pixels show topography and meteorology as the primary influences. However, these maps show which areas within the Blue Ridge area have which dependencies. For example, the southern tip of the region and the eastern foothills are primarily driven by meteorology while the high mountains are most affected by topography. However, as the summer progresses, these drivers reverse and reverse again. That is, in June and July, the high mountains begin to be meteorologically driven while August and September show a return to topographical control. The eastern foothills show a similar pattern with meteorology losing its importance during the middle of the summer but regaining it in September.

IV. Continental Scale Predictive Models of Vegetation

The next major step was to perform a similar analysis using GeoLearn to construct a dataset for the entire continental United States.

The data used were very similar to those used in the Blue Ridge analysis. The variable modeled was the EVI for the month of June. The explanatory variables were the same as those used in the Blue Ridge Region with the exception of the land cover categories, which were necessarily different due to the wider variety present across the continent. The land cover categories were: evergreen needleleaf forests, evergreen broadleaf forests, deciduous broadleaf forests, mixed forests, woodlands, wooded grasslands/shrublands, closed bushlands or shrublands, open shrubland, grasslands, and croplands. The data were divided into training (60%), and validation (40%) sets.

The training set was used to train regression trees up to 20 levels and an ordinary least squares (OLS) linear model. Using the estimated models, the summed squared error for both subsets of the data were computed. For comparison, the errors were transformed into the root mean squared error to provide the typical error standard deviation around the predicted values. This provides a summary measure of performance for the models. These values are shown in Fig. 5 with the OLS results plotted as horizontal lines for comparison with the tree results.

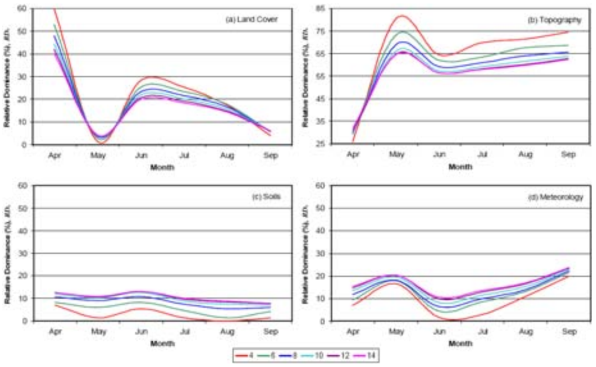


Fig. 3. Graphs showing how Amandian relative global dominance changes between months for land cover, topography, soils, and meteorology. The colors represent different depths of tree: red = 4 levels, magenta = 14. From [1].

The summary statistics reveal a few interesting things. The OLS training and validation results and virtually identical and 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 reaching a minimum at 17 levels. As context, the EVI in this dataset range from 0.1 to 1.0 with a sample mean of about 0.28 and a sample standard deviation of 0.145 .

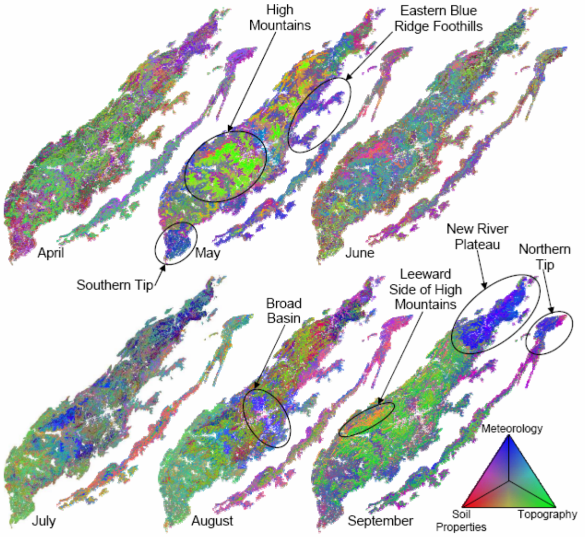


Fig. 4. Maps showing which type of explanatory variable is most influential at each location. From [1].

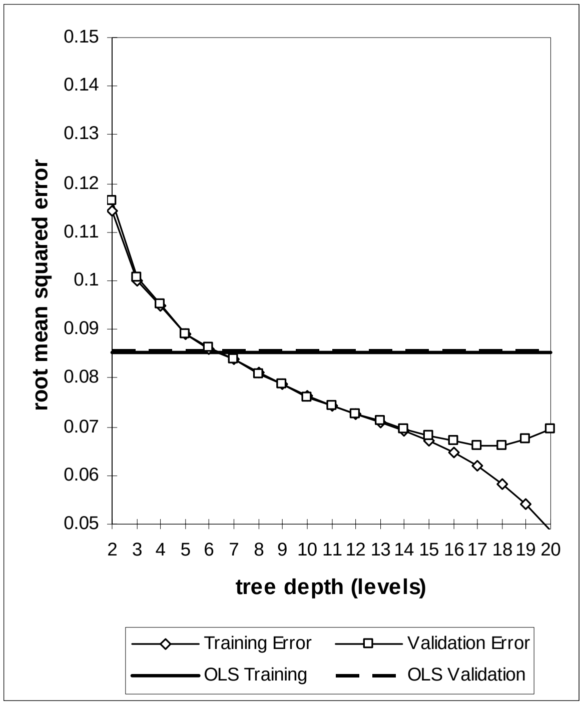


Fig. 5. Summary performance of regression trees and linear model.

Examination of the trained models can provide insight into the most important influences on the EVI at the continental scale. For the linear 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 (called β). To avoid scale of measurement effects, the coefficients were rescaled by the sample standard deviation of their corresponding explanatory variables. However, large values for these rescaled coefficients can still be misleading if the data are not able to support a good estimate. Employing the statistical assumptions of the model, the uncertainty of estimation for each coefficient can be computed. The coefficients and the *t*-statistics measuring their likelihood of being different from zero are plotted in Fig. 6. Due to the large amount of data, almost all the parameters are "statistically significant," but some are still clearly more certain than others. Note that the land cover was included as indicator variables for the different categories. The indicator for barren was excluded, and hence the estimated coefficients show the effect of the stated land cover *as compared to* the omitted category (barren). Hence, the land use coefficients are, for the most part, quite different from zero (indicating that they have a different effect than "barren"), but quite similar to each other. Considering both the magnitudes and the uncertainties, the most important variables seem to be: longwave radiation, shortwave radiation, precipitation, daytime temperature, nighttime temperature, soil pH, and possibly elevation. However, the day and nighttime temperatures have very nearly opposite coefficients. This could be due to a close linear correlation between the two. It is quite possible that they are so similar that only one should be included.

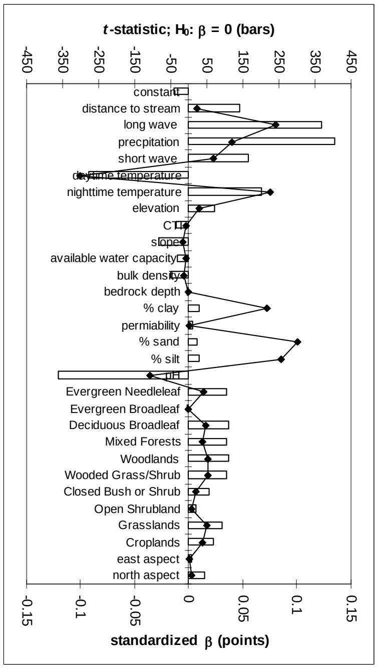


Fig. 6. Coefficient estimates and *t*-statistics for linear model.

The regression trees can yield similar information using the approach discussed with the Blue Ridge analysis. Using the 17 level tree, the Amandian Relative Global Dominance is computed for each variable as shown in Fig. 7. The most influential variable is longwave radiation followed by precipitation, soil pH, shortwave radiation, nighttime temperature, elevation, and the indicator for open shrubland. The regression tree identifies almost the same set of important variables as the linear model. Notice that the regression tree only picks nighttime temperature but not daytime temperature. Regression trees are much less susceptible to multicollinearity problems than OLS regression so this is a further indication that only one of the two temperatures is actually meaningful.

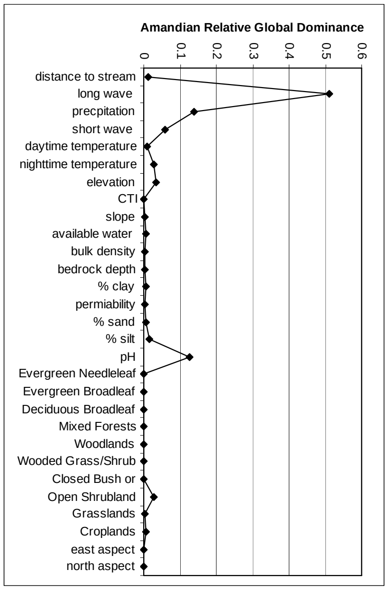


Fig. 7. Amandian Relative Global Dominance scores for 17 level tree.

V. Continental Scale Cluster Analysis

Another approach to understanding the relationships between remotely sensed variables is not predictive, but associative.

Using the tools in GeoLearn, continental scale datasets at one kilometer resolution are constructed for the following variables in the month of March: the Enhanced Vegetation Index (EVI), the Normalized Difference Vegetation Index (NDVI), the fraction of photosynthetically active radiation (FPAR), the leaf area index (LAI), emissivity, land surface temperature (LST), and albedo. The EVI, NDVI, and LAI are constructed measurements intended to reflect attributes of the plants being detected. The remaining variables measure various aspects of the incoming and outgoing radiation for each location which are thought to influence how well the plants grow. After cleaning, the final dataset consisted of about 5 million pixels.

The data were analyzed using a cluster algorithm based on a modification of the standard *k*-means clustering approach. A mean-squared-error measure is used to assess how well the clusters are defined with the goal being to find the locations of the specified (*k*) number of clusters that minimize this summary error. The entire dataset was randomly broken into a training set (70% of the original data) and a validation set (the remaining 30%) to help assess whether the discovered clusters are spurious or reasonable. The training set is used to discover the clusters. After they have been defined, the validation data are assigned to the newly minted clusters. If there is a great discrepancy in the error for the training set and error for the validation set, then the clusters are not likely to be correct or useful.

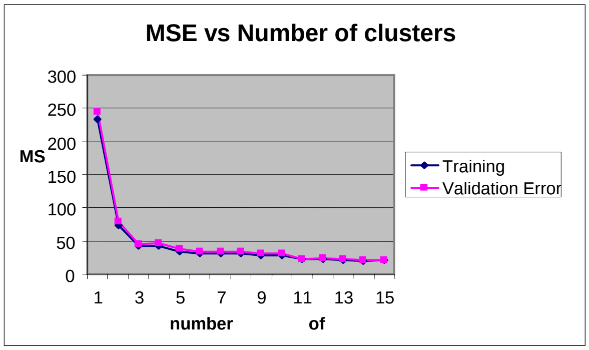


Fig. 8. Summary performance of different numbers of clusters.

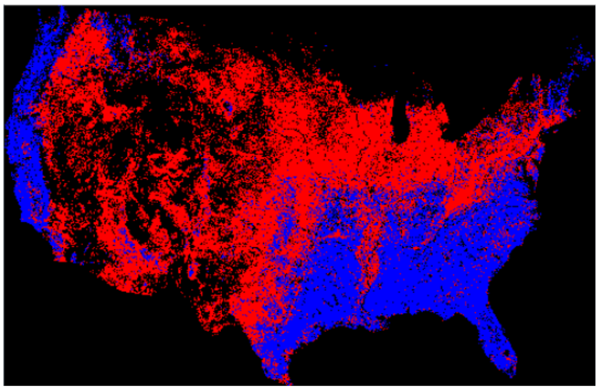


Fig. 9. Locations of clusters for two clusters.

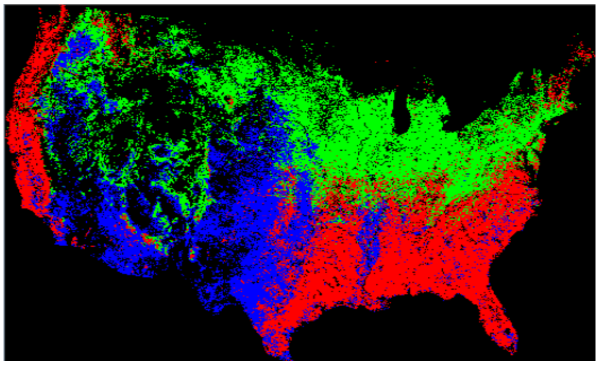


Fig. 10. Locations of clusters for three clusters.

Once clusters have been identified, each pixel can be annotated with which cluster it belongs to, resulting in a map showing the geographic locations of the members of each cluster (which are based on solely non-geographic data).

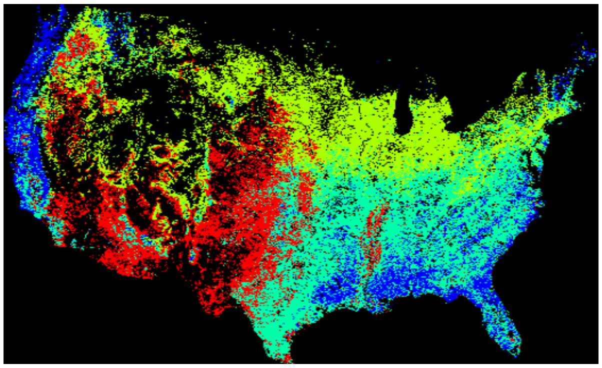


Fig. 11. Locations of clusters for four clusters.

Fig. 8 shows how the error decreases as the number of clusters is increased. The major improvements are realized when forming two and three clusters. Looking for more clusters improves the performance, but not in a very convincing way.

Fig. 9, Fig. 10, and Fig. 11 show where these clusters are located across North America. It is interesting that the clusters seem to have fairly distinct geographical boundaries even though geographic information is not provided in the analysis of the clusters.. The first two clusters (Fig. 9) show up geographically with the Midwest and Mountain West along with a sliver along the Mississippi River in the red cluster and the Pacific Coast grouped with the South and a bit of the extreme northeast of New England comprising the blue cluster. When three clusters were generated (Fig. 10), geographically, the same regions appear: the blue cluster from the two-cluster-regime is virtually the same as the red cluster in the three-cluster-regime. The old red cluster is split basically north/south. The four-cluster regime (Fig. 11) again shows the same regions with the original blue cluster being split this time into roughly coastal and interior regions.

VI. Conclusions

This project has expended considerable effort to build tools able to handle the large amount of remotely sensed data that has been largely untapped for scientific analysis beyond descriptive techniques. This paper has shown two approaches (predictive and clustering) to analyzing data related to regional and continental scale influences of vegetation growth.

Acknowledgement

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) under Award Nos. ESSF/O2-0000-0216 and NNG04GP78G, and the National Science Foundation (NSF) under Award Nos. EAR 02-08009 and EAR 04-12859. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NASA or NSF. We would also like to thank the National Center for Supercomputing Applications (NCSA) for their support through the D2K data mining platform.

References

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