Unraveling the Dominant Influences on the Evolution of Land-Surface Variables using Data Mining

Praveen Kumar[1], Peter Bajcsy[2], Amanda B. White[1], Vikas Mehra[1], David Tcheng[2], David Clutter[2], Wei-Wen Feng[2], Pratyush Sinha[1], and Richard Robertson[1]
[1]Department of Civil and Environmental Engineering,
[2]National Center of Supercomputing Applications,
University of Illinois at Urbana-Champaign,
Urbana, Illinois 61801
[contact e-mail:kumar1@uiuc.edu]

Introduction:

The objective of our research project is to develop data mining and knowledge discovery in databases (KDD) techniques, using the "Data to Knowledge" (D2K) platform developed by National Center for Supercomputing Application (NCSA), to facilitate analysis, visualization and modeling of land-surface variables obtained from the TERRA and AQUA platforms in support of climate and weather applications.

The project targets to address the science question: "How is the global Earth system changing?" In particular it focuses on the theme: What factors influence/modulate the changes in global ecosystem? The specific science questions that this project is focused on are:

1) How are evolving surface variables such as vegetation indices, temperature, and emissivity, as obtained from the TERRA and AQUA platforms, dynamically linked?
2) How do they evolve in response to climate variability such as ENSO (El Niño Southern Oscillation)? and
3) How are they dependent on temporally invariant factors such as topography (and derived variables such as slope, aspect, nearness to streams), soil characteristics, land cover classification, etc?

Answers to these questions, at the continental to global scales will enable us to develop better parameterization of the relevant processes in forecast models for weather, and inter-seasonal to inter-annual climate prediction. However, answering these questions at the continental to global scale requires the ability to perform analysis of a multitude of variables using very large datasets. Our data mining system is building this capability for Earth science datasets being collected by NASA.

Data Mining System:

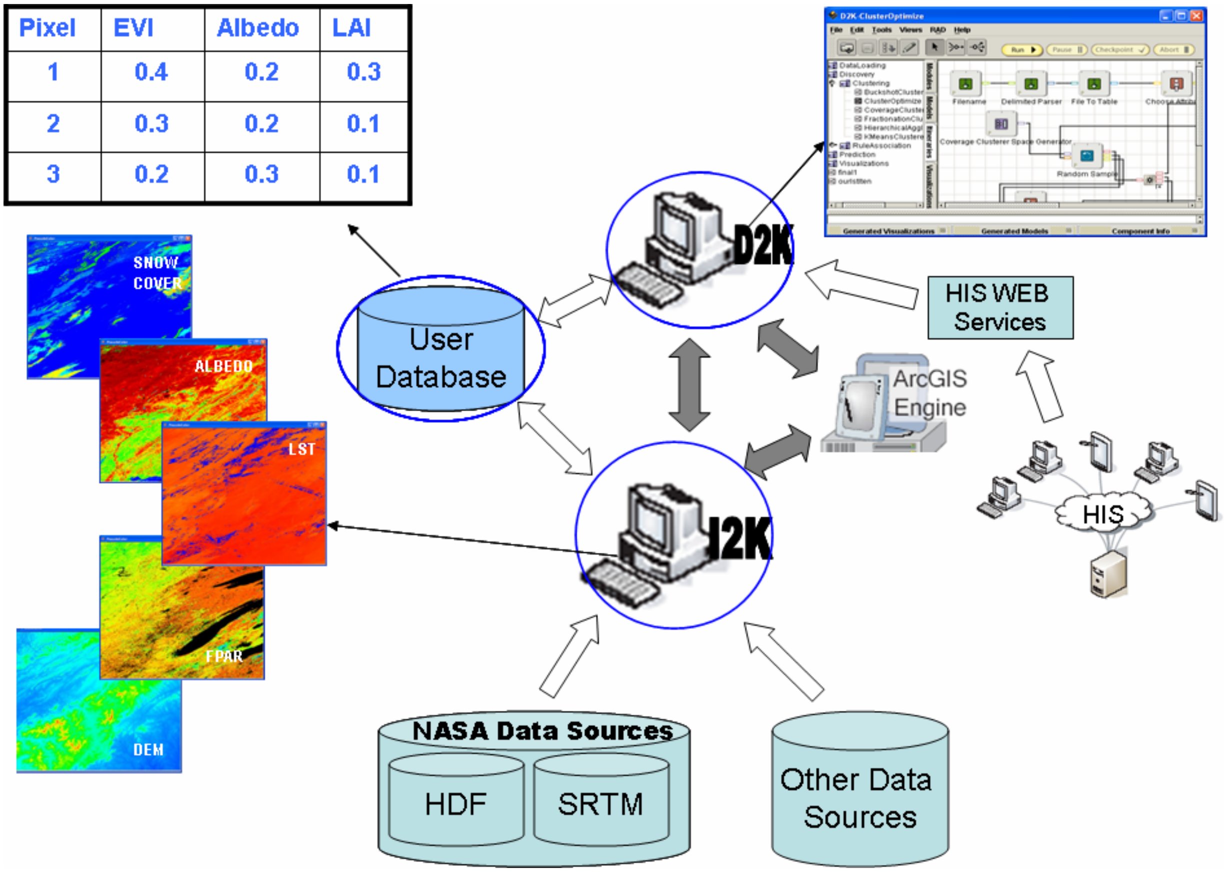
To support various data formats, a common interface is designed to visualize, preprocess and analyze the data. Some of the supported data formats include hierarchical data formats (HDF), digital elevation model (DEM), and geographical information system (GIS) supported vector files. The overall system architecture has been divided into four parts (Fig. 1). These components are explained below:

1) *Read raster data using I2K:* I2K is an image analysis tool, designed to automate processing of huge datasets and is capable of analyzing multi-dimensional and multivariate image data. When analyzing multiple geographic datasets over the same geographic area, it is necessary to preprocess and integrate heterogeneous datasets. I2K is a key component for preprocessing, visualizing and integrating the diverse datasets. I2K uses HDF libraries to load HDF data, and links to ArcGIS Engine functionalities to operate on GIS data formats. Fig. 1

shows the visualization of different scientific data sets: Snow cover, Albedo, LST (Land Surface Temperature), FPAR (fraction of Photosynthetic active radiation) and DEM.

2) *ArcGIS Engine*: It is a complete library of GIS components which can be embedded into custom applications. I2K links to these libraries for features extractions e.g. calculation of slope, aspect, and flow accumulation grid from DEM. These derived variables are used for analysis along with the remote sensing data sets.

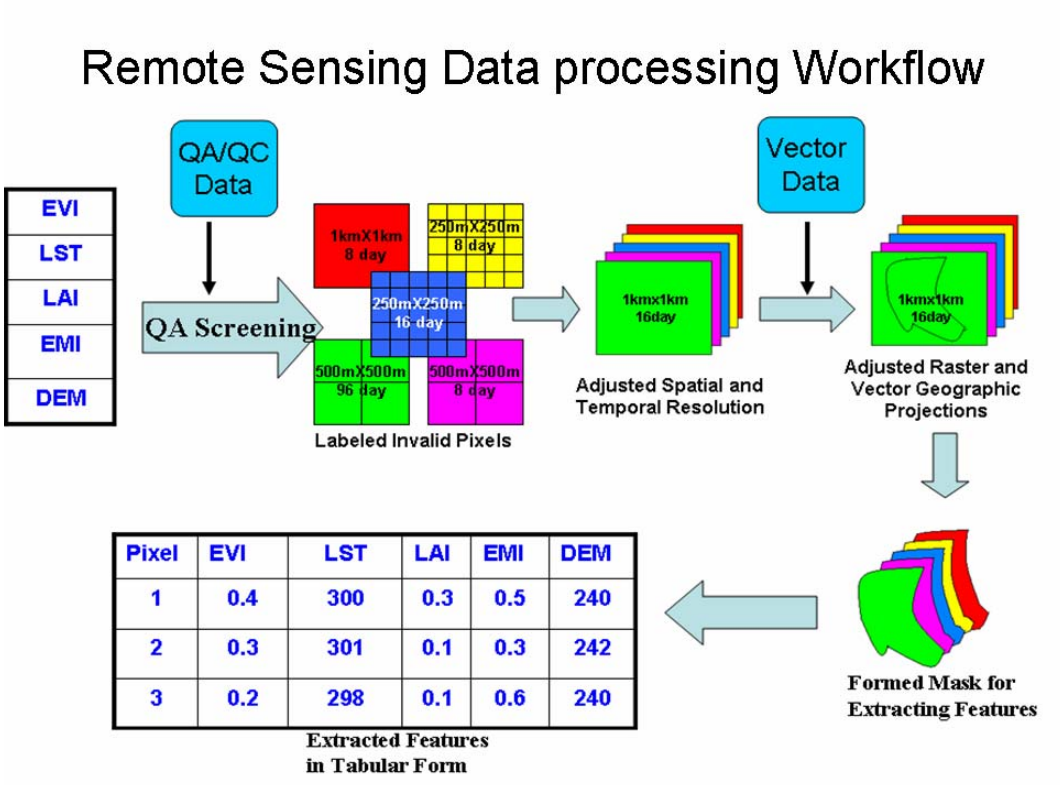
3) *Create Relational Database:* Creating a user Database (Fig. 2) is a data preprocessing and integration step. Different scientific datasets such as Enhanced Vegetation Index (EVI), Albedo, Leaf Area Index (LAI), Emissivity and Sea Surface Temperature (SST) are at different spatial and temporal resolution. Also there is quality assurance and quality control (QA/QC) data associated with each scientific variable. QA/QC data provide information about the quality of data for each pixel inside a scientific dataset.



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

Figure 1: Illustration of overall system architecture for data ingestion, preprocessing, integration, visualization and data analysis using various data mining algorithms. I2K reads all data sets from different data sources and visualize (snow cover, Albedo). It calls GIS functions using ArcGIS engine interface to perform feature extraction tasks (slope, aspects). All measured and derived variable are ingested in to database after preprocessing (spatial and temporal adjustment, removing bad pixels using QA/QC). D2K is used to analyze this database and results are visualized in I2K. D2K can also ingest data through web services such as those of Hydrologic Information Science program of CUAHSI.

To create an analysis database, we need to choose a unique spatial and temporal resolution. This is done by upscaling or downscaling the data. The unique spatial and temporal resolution is supplied by user as an input before creating the database. User may be interested in analyzing the data for a particular region only (Fig. 2). In that case (s)he can create a mask by selecting the area that (s)he wants to analyze. QA/QC data is used to remove bad pixel values e.g. no data values or bad pixel data received by satellite due to clouds. This option is again provided by the user. After all the above processing is done, integrated scientific and derived data sets are written into a database (Fig. 2).


Remote Sensing Data processing Workflow
QA/QC Data
Vector Data
EVI
LST
LAI
EMI
DEM
QA Screening
1kmX1km 8 day
250mX250m 8 day
250mX250m 16 day
500mX500m 96 day
500mX500m 8 day
Labeled Invalid Pixels
1kmx1km 16day
Adjusted Spatial and Temporal Resolution
1kmx1km 16day
Adjusted Raster and Vector Geographic Projections
Formed Mask for Extracting Features
Extracted Features in Tabular Form

Pixel	EVI	LST	LAI	EMI	DEM
1	0.4	300	0.3	0.5	240
2	0.3	301	0.1	0.3	242
3	0.2	298	0.1	0.6	240

Figure 2: Remote sensing data processing workflow.

4) *Use D2K for data mining:* This task plays the central role to enable automatic knowledge discovery through data mining. D2K uses database created in the above step as an input. It has modules for a variety of algorithms like multiple regression, Naïve Bayes, Decision Tree, and Neural Network to find various characteristic of data sets. Scientific question which we aim to answer are: (1) identify the dependence of the dynamically evolving variables on each other and their temporal scales of variability; and identify the roles of climate variability as a determinant of the variability in the dynamically observed quantities (2) identify how land-surface characteristics (elevation, slope, aspects, soil properties etc) further modulate the dynamical evolution of vegetation.

Results:

Initial study is performed over the Blue Ridge Ecoregion (CEC/EPA Level III) in the Appalachian Mountains, and the objective is to identify and analyze the dominant influences on the remotely-sensed vegetation greenness throughout the growing season. Regression tree induction is employed and a relevance index is developed based on the generated regression trees to examine the spatio-temporal variability of the controls on the vegetation growth. Within the ecoregion, cohesive areas are found where the vegetation dynamics differ from the surrounding areas, and in comparing these with the next-smaller scale ecoregions (EPA Level IV), similarities and differences are observed. Overall, the dominant control on the vegetation growth is the topography. The second-most influential controls are the meteorology and land cover, though at different times during the growing season, and the least influential is soil properties.

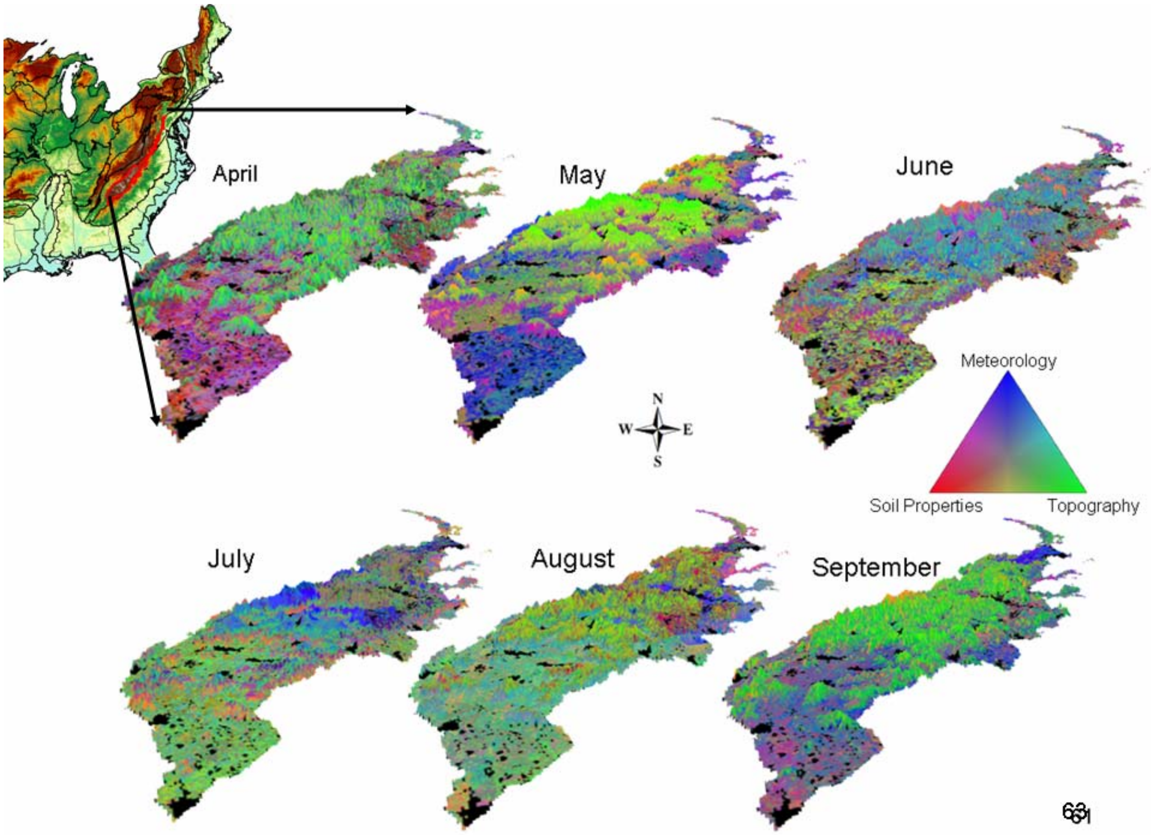


Figure 3: This figure illustrates the factors that govern the growth pattern of EVI (enhanced vegetation index) during the month of May through September over the Blue Ridge ecoregion. Meteorological factors include precipitation, long and short wave radiation, and day and night time temperature. Topographic factors include elevation, slope, aspect, distance to nearest stream, and topographic index. Soil properties include percent sand/silt/clay, permeability, available water capacity, depth to bedrock, and pH. The analysis is performed using a decision tree algorithm. [Adapted from White and Kumar, 2006.]

Summary:

We have developed a data mining system that is capable of handling very large remote sensing data, along with other raster data. A variety of functions are implemented to ease the task of preparing these data for the execution of mining algorithms. Analysis over the Blue Ridge Ecoregion has been a pilot study that has informed and guided the development of the system, and is available to public at large through the following web link; http://cee.uiuc.edu/research/hydrology/hydroinf_Intro.html.

References:

White and Kumar (2006): Dominant Influences of Vegetation Greenness. (a) Part I: The Blue Ridge Ecoregion. Submitted to Journal of Geophysical Research – Biogeosciences.

Acknowledgement: NASA grants NNG04GP78G and ESSF/O2-0000-0216 and NSF Grant EAR 04-12859.