Discovery of Geospatial Discriminating Patterns From Remote Sensing Datasets

Josue Salazar

Mentors:
Dr. Tomasz Stepinski
Dr. Wei Ding

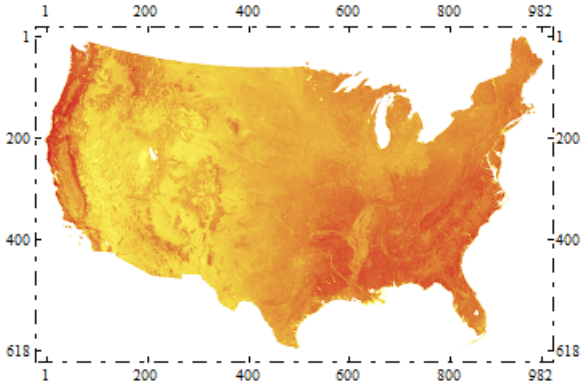
REU UHD Program Summer 2009

Overview

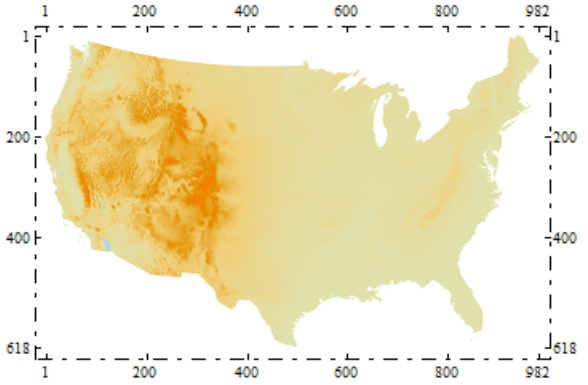
- Introduction
- Project objective
- Following previous work
- Summary of the method
- Conclusion

Introduction

- Remote sensing geospatial observations
  - Multispectral Images
  - Surface topography
- Environmental quantities
  - Mineralogy
  - Land cover
  - Soil properties
  - Vegetation density
  - Surface temperature
  - Precipitation



Class Variable: Average Vegetation



Explanatory Variable: Elevation

Project Objective

- Develop and apply a tool which auto-analyzes a set of class and explanatory geospatial variables that provides a complete description of the class variable dependence on the explanatory variables.
- Techniques used:
  - Association analysis
  - Reinforcement learning
  - Similarity measurement

Following previous work

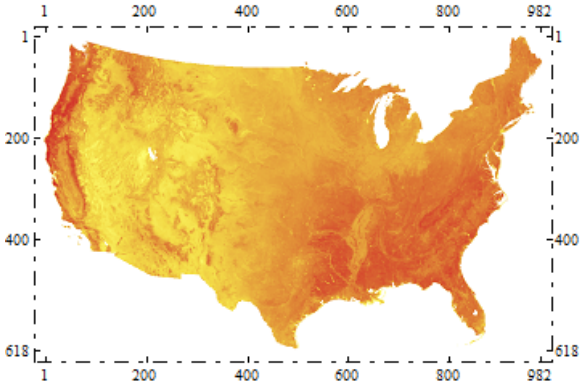
- Work done:
  - Basic design of the tool was research and implemented.
  - Tool was applied to an example involving vegetation density within the US (class variable) and 9 environmental variables (explanatory variables).

- Improvements:
  - Code design
  - Improvements in calculating optimal region of interest and the optimal region of discriminating patterns.
  - Similarity measure between patterns of explanatory variables.
  - Tool is being applied to a new dataset pertaining to Mars. The outcome will be a complete description of thermal inertia of Martian surface (class variable) in terms of up to 19 explanatory variables including surface roughness, mineral composition, and surface brightness.

Summary of the method

1. Discretization of continuous datasets.
2. Mapping datasets to one super dataset.
3. Mining for discriminating patterns.
4. Boundary optimization
5. Pattern Sumarization
6. Description of class variable dependence on explanatory variables.

1. Discretization of continuous datasets.

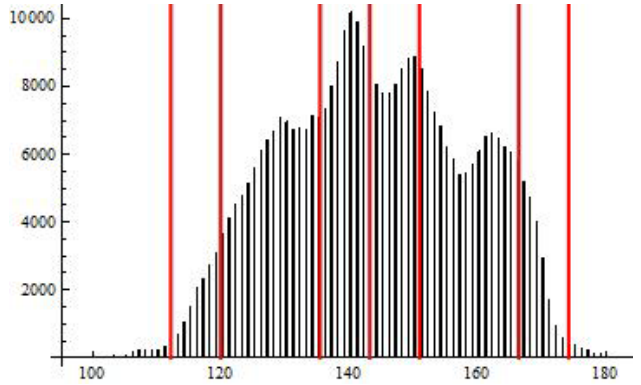


Continuous Average Vegetation

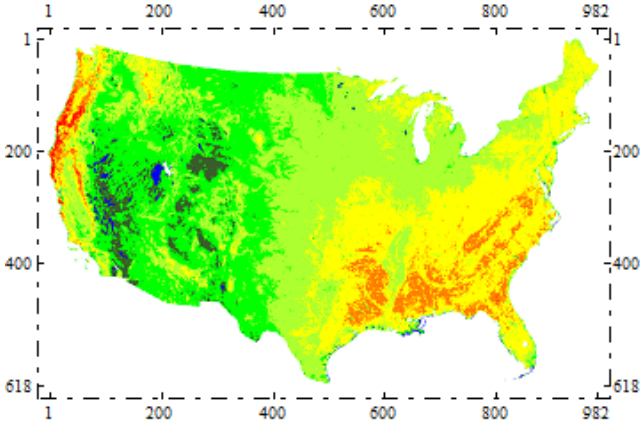
aveveg:

FileName	Mean	TrMean	Median	SD	MAD	SN
aveveg	143.804	143.777	143.	14.589	11.	15.5038

split1: 111.992 split2: 119.744 split3: 135.248 split4: 150.752 split5: 166.256 split6: 174.008

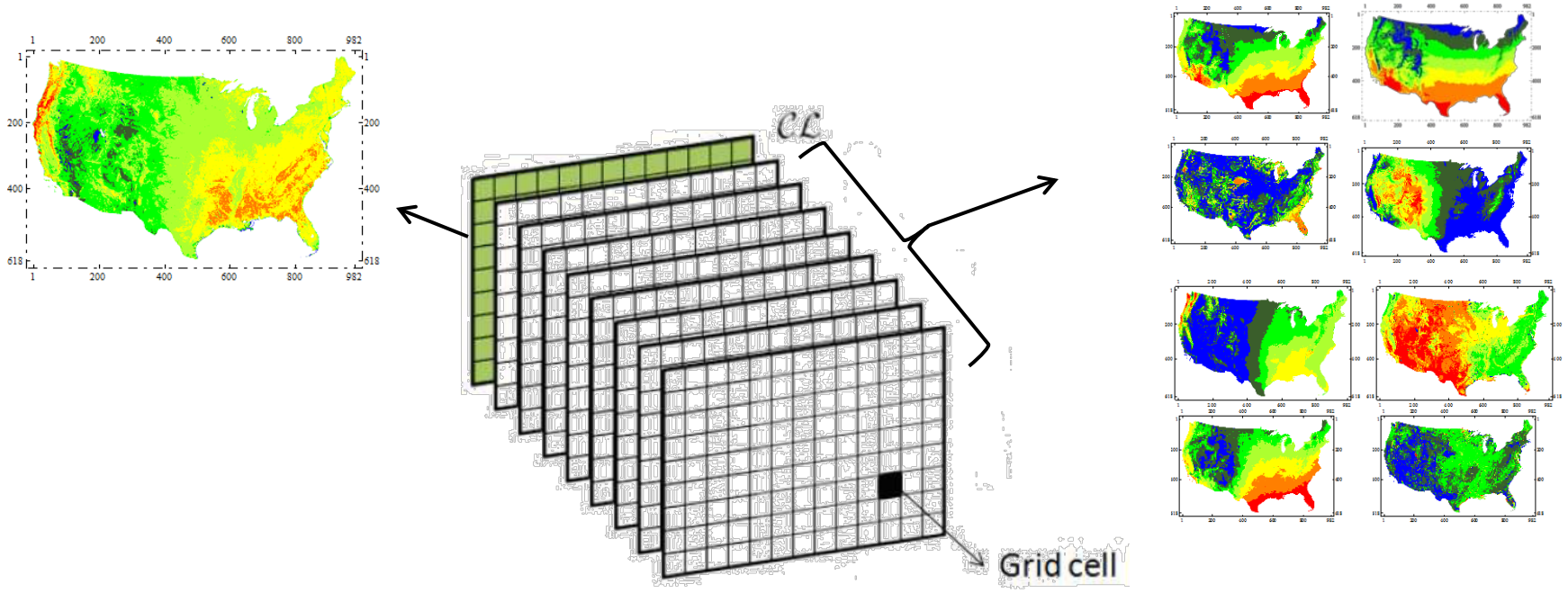


Average Vegetation Histogram



Discrete Average Vegetation

2. Mapping datasets to one super dataset.



Each cell of the super dataset contains an object vector with the values of all datasets on the same coordinate.
Example:

{1,4,6,2,2,3,4,7,7 }

^Class variable value

3. Mining for discriminating patterns.

- What are the patterns of the explanatory variables that explain the distribution of high vegetation density in the US?
- To find these patterns we use the concept of association analysis, in particular, closed patterns and emerging patterns.
- The result of the application of association analysis is an effective minimal representation of the high vegetation region supported by a set of explanatory variable patterns.
- Example:

Object → {1,4,6,2,2,3,4,7,7 }

^Class variable high value

Patterns supporting object (vector)

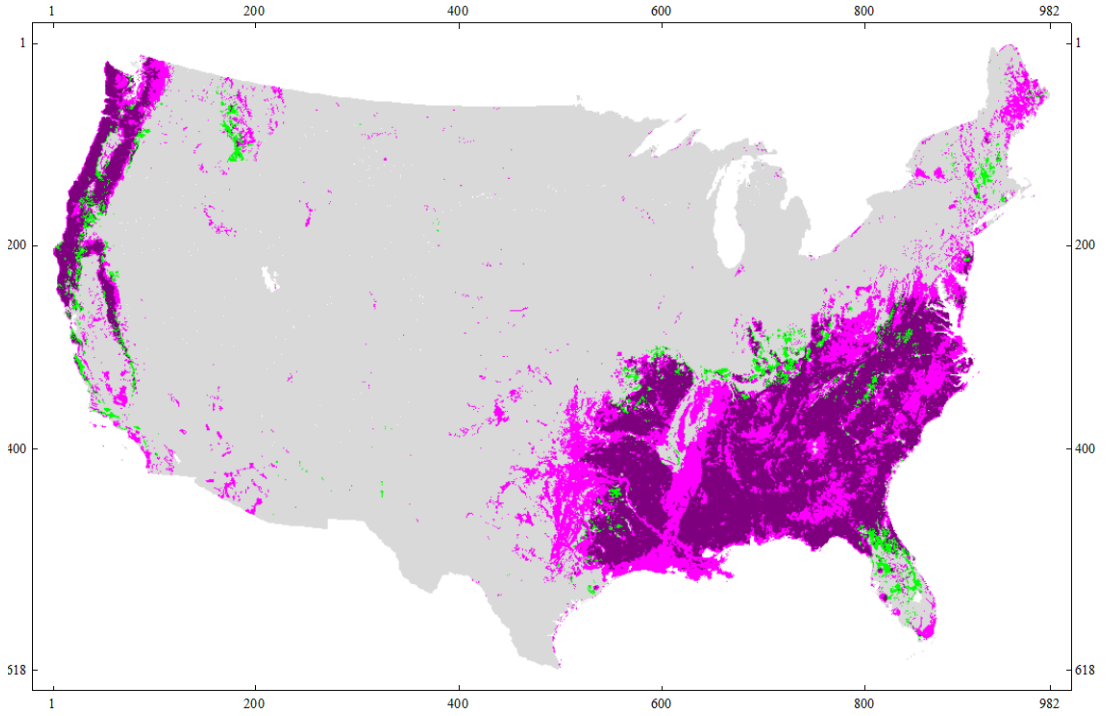
{1,_,_,2,2,3,_,_,7}
{_,4,_,_,2,_,_,_,7}
{_,_,6,_,_,_,4,7,7}

^Class variable high value

3. Mining for discriminating patterns.

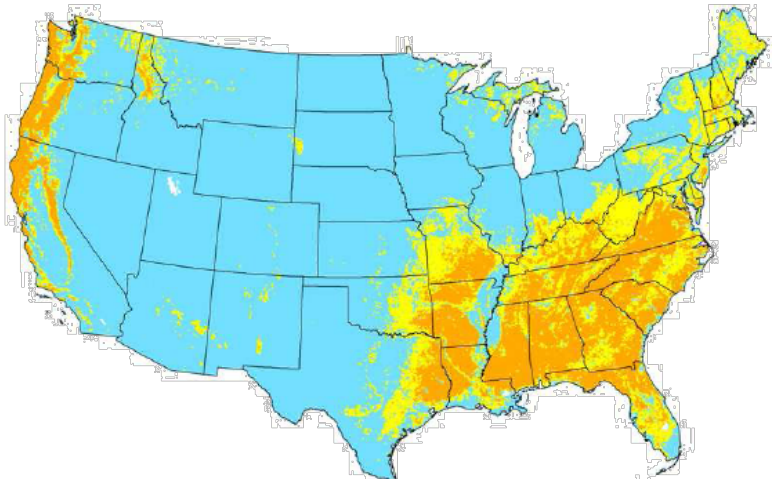
- Once we have the set of closed patterns, only those patterns with the most frequency on high vegetation emerge.
- The other patterns are disregarded.

Green – HV
Magenta – Footprint of emerging closed patterns
Purple – Intersection

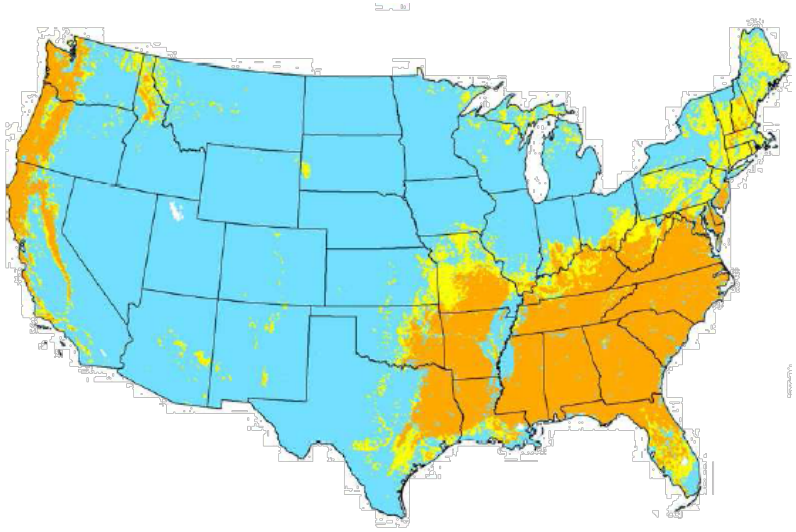


4. Boundary optimization

- Original datasets contain real continuous values.
- Selecting a threshold value is an arbitrary choice.
- This creates artificial sharp boundaries and leads to loss of information.
- We utilized the reinforcement learning method to smooth the transition from high to low values.



HV – 7
LV – 6



Sharp boundary between high vegetation (orange)
And not-high vegetation (yellow)

Smoothed boundary high vegetation boundary (orange)

5. Pattern Summarization

- As result of the previous two processes we have a set of k number of explanatory variable patterns which can be summarized by clustering them into a small number of “super patterns.” This will allow the domain scientist to analyze the controlling factors of the class variable.
- In order to cluster the patterns, a distance measure had to be defined.
- Distance=(1/Similarity)-1
- The similarity between pattern X and Y is defined by the following definition.

$$s(X,Y) = \frac{\sum_{i=1}^{m} s(X_i, Y_i)}{m}$$

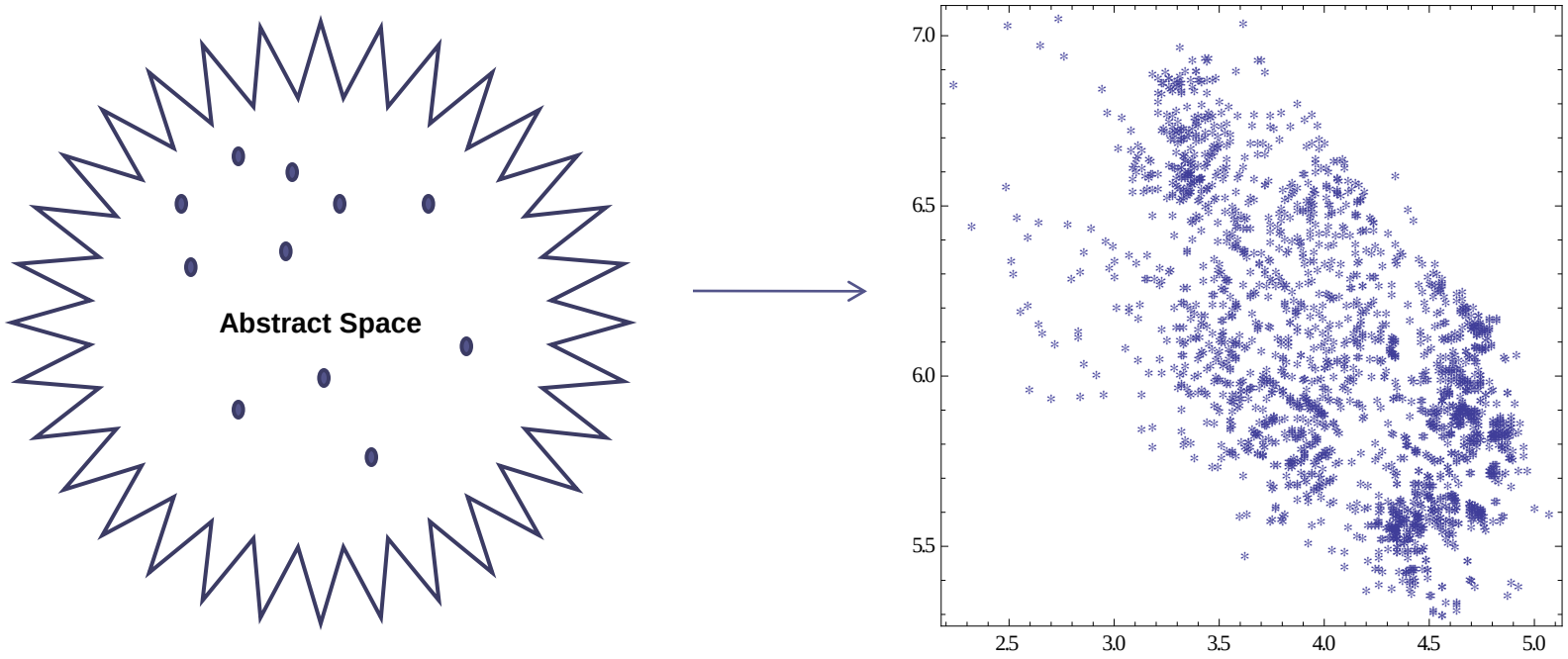
- In summary, the patterns are aligned, calculate the similarity between each feature and then take the mean of the m similarity values as the overall similarity between the patterns.

Pattern X: {3,_,5,_,}
Pattern Y: {6,1,_,_}

S(X,Y)=[S(3,6)+S(_,1)+S(5,_)+S(_,_)]/4

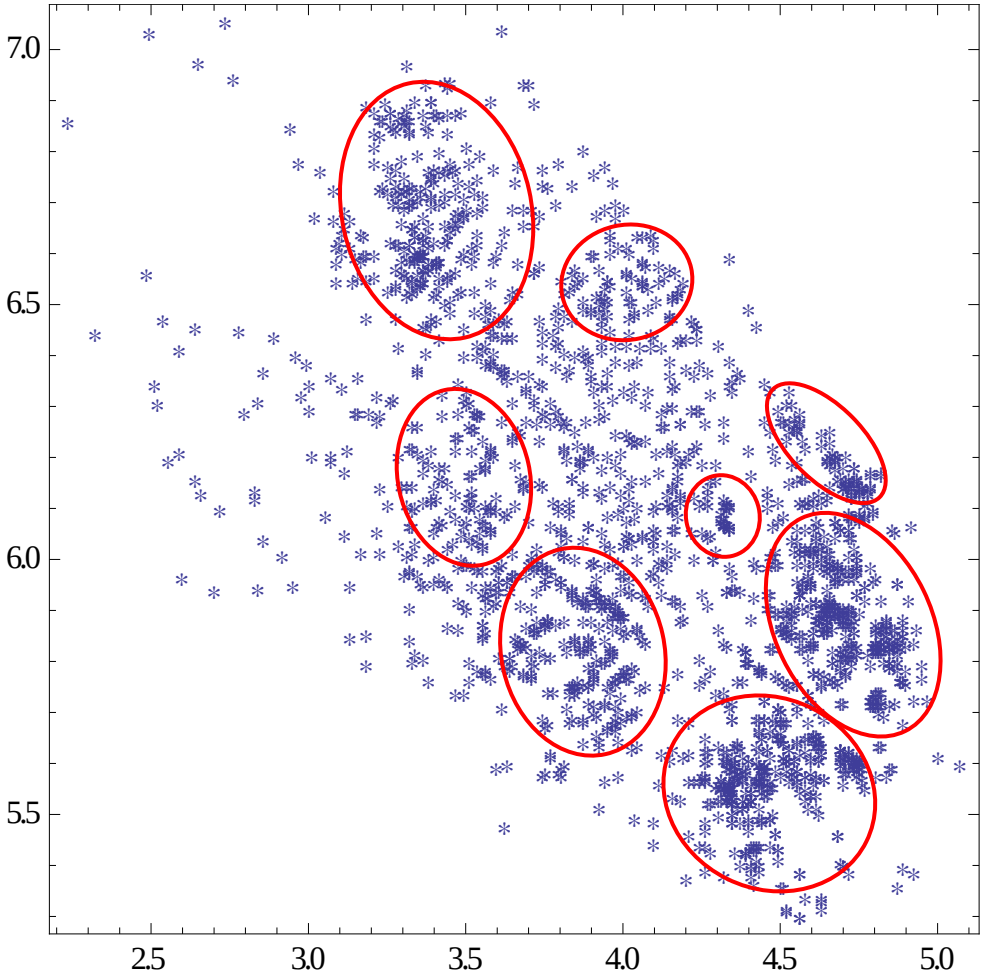
5. Pattern Summarization

- Using the distance measure we calculate the distance between all the emerging patterns to use it as input to the Sammon's Mapping algorithm.
- The Sammon's mapping algorithm creates a two dimensional representation of the distances between the patterns in the abstract space.



5. Pattern Summarization

Work in progress…



6. Description of class variable dependence on explanatory variables.

- The footprint of each “super pattern” (cluster) found using the Sammon’s map is graphed together with the class variable.
- By looking at the “super patterns” special characteristics, the domain scientist can determine what are the factors of high vegetation and gain knowledge of the reasons its distribution.
- We are hoping that the footprint of each “super pattern” doesn’t overlap with the others.
- This would indicate that there exits a high amount of vegetation on different regions of the US due to different reasons.

Conclusion

- Further refining work needs to be done for patterns summarization in order to gain more understanding of our method.
- The discovery of geospatial discriminating patterns method is being applied to the US and Mars datasets.
- The pattern summarization method is being applied to other non-spatial datasets.