REMOTE SENSING FOR ENVIRONMENTAL BASELINING AND MONITORING

D. B. Reister

Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6016

J. K. Otton
U.S. Geological Survey

J. C. Ritchie
U.S. Department of Agriculture
Agricultural Research Service

ABSTRACT

Hyperspectral remote sensing is an emerging technology with the potential to provide detailed environmental data for a region, rapidly and at low cost. In October 2003, hyperspectral images were acquired in Osage County, OK, a major oil production area. ORNL has developed clustering methods for grouping hyperspectral measurements into as many distinct classes as possible and applied them to the Osage County measurements. Clusters have been found for two cluster radii. For the larger radius, the number of clusters ranges from 35 to 133 and the range for the smaller radius is from 100 to 418.

INTRODUCTION

Hyperspectral remote sensing is an emerging technology with the potential to identify plant species, map vegetation, characterize soil properties, identify contamination, classify ecological units and habitat characteristics, and differentiate causes of vegetation stress. The main benefit of this project, both to the DOE Fossil Energy Program and to the petroleum industry, will be an improved ability to provide detailed environmental data for a region, rapidly, at low cost. Increasingly, environmental data are needed to assess present conditions of lands owned, leased or managed by petroleum companies, and to characterize and quantify changes in the environmental conditions of these lands through time. Present methods of assessing large areas depend extensively on field surveys, which can take weeks or months to complete. Such methods are inconvenient and can be expensive. Further, some areas that are inaccessible or very large are difficult or virtually impossible to monitor accurately using field-survey techniques only. Hyperspectral sensing can be used to: identify plant species, map vegetation, characterize soil properties, identify contamination, classify ecological units and habitat characteristics, and differentiate causes of vegetation stress. The petroleum industry needs reliable analysis tools for rapid quantitative measurements and identification of type of stress causing environmental impact.

The oil industry is responsible for large tracts of land all over the world, used for production, refining, or marketing. If this project is successful at developing appropriate algorithms for the oil industry, remote sensing offers tremendous opportunities for the industry. Although many algorithms have been and are being developed, their focus is primarily on resource management (forest fires) and agriculture (crop nutrient and water needs). Algorithms needed by the oil industry (identification of hydrocarbons and differentiation of different types of stress relevant to our operations) are not currently available. With

remote sensing, there is the ability to reduce the time to acquire important environmental information from months to days. Further, the data acquired can have a resolution of meters as opposed to current field monitoring, where sampling sites often are tens of meters apart. Remote data can be collected on a regular basis through time to detect and quantify environmental change. Regular data collection has a huge impact on the ability of the oil industry to baseline and monitor environmental issues. By understanding environmental conditions, capital and remediation projects can be accelerated, saving millions of dollars. By monitoring conditions more closely and in a timelier manner, we can improve management and protection of the environment.

An approximate estimate of the cost of current field monitoring is $20,000/acre. The cost of HyVista hyperspectral imagery is about $60 K for 80,000 acres (324 square km). Using commercial software (ENVI), preliminary analysis of the measurements to produce geobotanical maps costs $10 K. Thus, after R&D is completed, the total cost of producing geobotanical maps may be about $1/acre. Advanced methods using satellite data may be able to reduce the cost of hyperspectral imagery component of the total cost.

Produced water is the largest volume waste product in upstream activities. The oil industry needs low cost accurate methods to quickly detect environmental discharges of produced water in oil fields that they operate or plan to purchase. DOE is supporting this project because it is developing an important environmental monitoring technique that does not currently exist.

DISCUSSION OF CURRENT ACTIVITIES

SITE SELECTION

The US Geological Survey (USGS), the USDA Agricultural Research Service (ARS), and ORNL are collaborating to develop remote (airplane or satellite) hyperspectral sensor techniques to identify areas impacted by oil production. This is the final year of a multiyear project. Initially the project was focused on an oil spill site at the Jornada Experimental Range near Las Cruces, NM. New methods were developed to analyze hyperspectral field data measurements that were collected at Jornada. This year our focus shifted to Osage county, OK, which is the Osage Indian Reservation and has been a major oil producing area (38,500 oil wells) since 1896. The county is large (2,260 square miles) and 1,480 square miles are within a quarter mile of an oil well. The Osage Nation owns the mineral rights and there are 500 independent operators in the county. Many areas in the county have brine scars or weathered oil pits (The USGS has sample photographs: http://ok.water.usgs.gov/skiatook/Skiatook. Photo.html). The objective of the project was to collect hyperspectral remote data in selected regions of Osage county and analyze the data to detect brine scars, oil pits, and plant stress associated with brine and oil.

In November 2002, the USGS and ORNL visited Osage County to select potential regions to image. The USGS has been conducting research at two sites (Site A and Site B) that are adjacent to Skiatook Lake and owned by the Army Corps of Engineers. We selected the 2 USGS sites, 6 sites on the John Zink Ranch, 4 sites on the Tallgrass Prairie Preserve (owned by the Nature Conservancy), and 5 sites on the Bluestem Ranch.

ORNL selected the HyVista Corporation (http://www.hyvista.com/) to measure the hyperspectral data. In May 2003, ORNL contacted HyVista and requested that they collect images of seven regions that total 39 square kilometers. One of the seven regions is shown in Figure 1. The red markers on the figure are Global Positioning System (GPS) waypoints that were measured during the field visit in November 2000.

The region in Figure 1 includes USGS Site B (114), and three Zink Ranch sites: 115, 124-125, and 126-127. On October 12, HyVista made the images. On October 27 and 28, a team from the USGS, ARS, and ORNL measured hyperspectral field data at 13 sites in the seven areas where the remote data was measured.

On November 10, ORNL received the data from HyVista . HyVista created a flight path that imaged three regions with overlapping strips. Each strip is 512 three-meter-pixels wide, for a total of 1536 meters. The North region has three strips with a total length of 12 km and a total area of 18 square km. The South region has 6 strips with a length of 69 km and an area of 106 square km. The West region has 5 strips with a total length of 40 km and an area of 61 square km. The total area is 185 square km. The data was received on five DVD and totals 14.4 gigabytes. The latitude and longitude of each pixel was provided. ORNL verified that the data covered our 7 study regions.

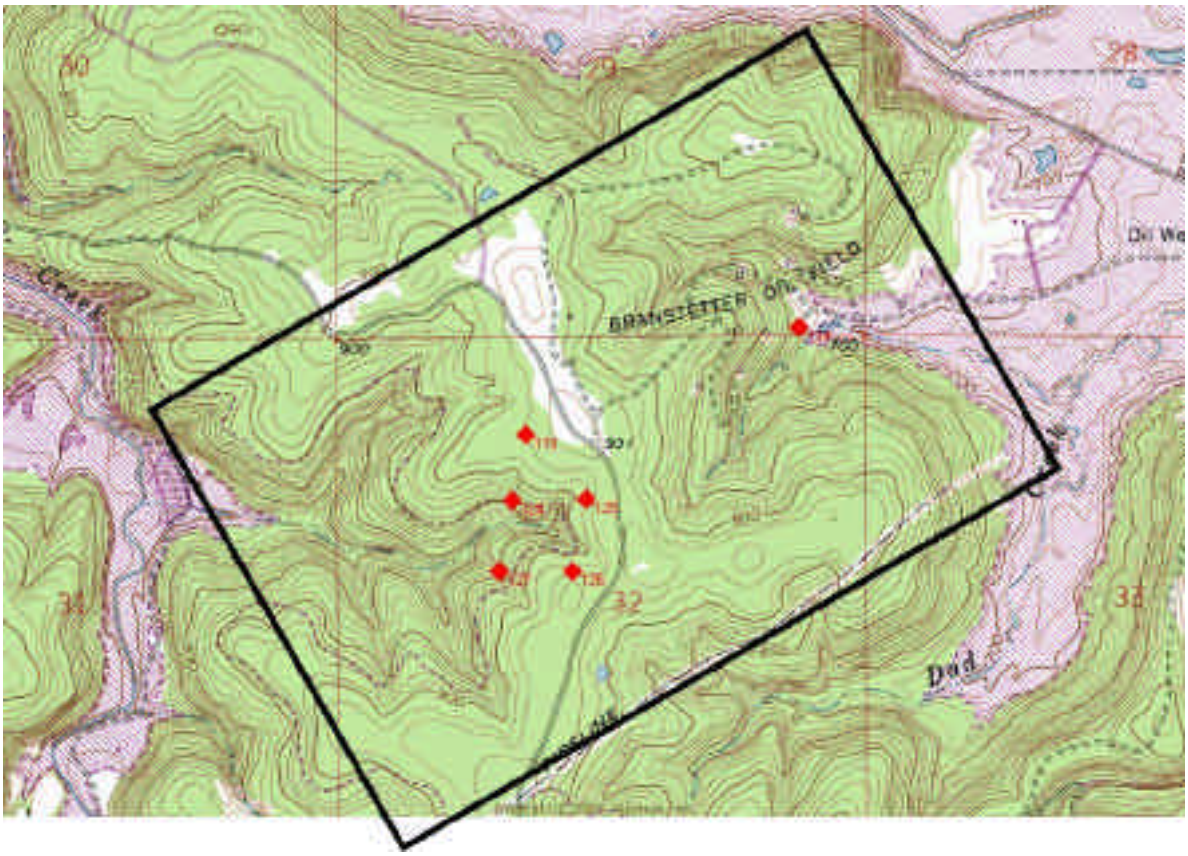

Figure 1. Selected region for Site B and three of the Zink Ranch sites.

The HyVista data has 126 spectral bands. Using three of the bands we can create the color image in Figure 2.

Figure 2. Three band HyVista image of Site B on the right side of Figure 1.

ANALYSIS

Our objective is to group hyperspectral measurements into as many distinct classes as possible. We can distinguish between information classes and spectral classes, where information classes are human classifications and spectral classes are groups or clusters of measurements made by a computer. A botanist uses many features to classify a plant in a particular information class: leaf type (flat, needle, deciduous, succulent), leaf geometry, attributes of branches, flowers and seeds, and smell. A hyperspectral reflectance measurement consists of hundreds of numbers that can range between zero and one. Spectral classes are groups of measurements with similar values for their reflectance spectra.

Hyperspectral sensors measure radiance and the data can be transformed to reflectance. For each of N spatial measurements, we transform the data by subtracting the mean of the M spectral values and normalizing to unit length. After the transformation, the data [a_{ij}, where $i = 1, M$ and $j = 1, N$] have the following properties: the spectral sum of data values is 0 and the spectral sum of the squares of the data values equals 1.

$$\sum_{\mathbf{i}} \mathbf{a_{ij}} = 0 \tag{1}$$

$$\sum_{\mathbf{i}} \mathbf{a_{ij}^2} = 1 \tag{2}$$

The components of the correlation matrix (c_{jk}) are the correlation coefficient between the spectrum at pixel *j* and the spectrum at pixel *k*:

$$\mathbf{c_{jk}} = \sum_{\mathbf{i}} \mathbf{a_{ij} a_{ik}} \quad (3)$$

For hyperspectral data, the correlation matrix can be huge (If the data has millions of pixels, the full correlation matrix will have trillions of elements). Ultimately, we need to calculate the correlation matrix between a few measurements that will be the root vectors for the clusters and all of the measurements (a matrix with 100 rows and 100,000 columns). Iterative methods can be used to screen the measured vectors and find candidates for the root vectors.

The Euclidean distance (d_{jk}) between two measurements is defined by:

$$\mathbf{d}_{\mathrm{jk}} = \sum_{\mathrm{i}} \left(\mathbf{a}_{\mathrm{ij}} - \mathbf{a}_{\mathrm{ik}}\right)^2 \quad (4)$$

Using (2) and (3), the Euclidean distance is directly related to the components of the correlation matrix

$$\mathbf{d_{jk}} = 2\left(1 - \mathbf{c_{jk}}\right) \quad (5)$$

We will define the cluster radius (r_c) by:

$$\mathbf{r_c} = \max\left[1 - \mathbf{c_{jk}}\right] \quad (6)$$

Our clustering algorithm depends on two parameters, the cluster radius (r_c) and the minimum number of members in a cluster (n_c). For each value of the cluster radius, we define the set of spectral neighbors ($\psi_{\mathbf{k}}$) for each spatial measurement point:

$$\psi_{\mathbf{k}} = \left\{\mathbf{j} \mid \mathbf{j} \quad [1, \mathbf{N}], \mathbf{c_{jk}} > 1 - \mathbf{r_c}\right\} \quad (7)$$

Since the limits on the correlation coefficient are ± 1, the cluster radius will be in the interval [0,2]. For a sufficiently small cluster radius, all of the neighborhoods ($\psi_{\mathbf{k}}$) will have one member (the point k). For a sufficiently large cluster radius, all of the neighborhoods will have all N points. When the cluster radius is given, no iterations are required to calculate the members of each neighborhood.

We start with an arbitrary root vector and calculate the correlation coefficients with all of the other vectors. We can count the members of the neighborhood. If there are more than n_c members, we have a cluster. To choose the next root vector, we look for a vector that is not close to any of the current root vectors. We need a criteria for "not close". Each cluster is a ball in an M dimensional space with a radius of r_c. A necessary condition for two spheres to be non-overlapping is that the distance between their root vectors must be at least 2 r_c. A necessary condition for many spheres to be non-overlapping is that the minimum distance between all possible root vectors must be at least 2 r_c. But we cannot know all possible root vectors when we start to choose root vector. We find that the clusters overlap and that the

overlap decreases when we define more root vectors and as we increase the distance from 2. Thus, our criteria for "not close" is that the distance between two root vectors must be more than $r_a = \alpha\, r_c$, where α is greater than 2 (currently, $\alpha = 2.8$). Another alternative would be to have value of α that is high for a small number of root vectors and decreases to 2 as the number of root vectors increases.

We are attracted to the terminology for human population patterns on the surface of the Earth. A cluster could be like a neighborhood in a large city or a subdivision in a rural county. We want to define distinct clusters, but can allow the clusters to overlap. We will classify each vector by its distance from the closest root vector. If the distance is less than r_c, the vector is a member of the cluster. We will define a group distance r_g that is analogous to the size of the city or county. If the distance is more than r_c and less than r_g , the vector is associated with the cluster. If a vector is farther than r_g from all of the root vectors it will be unclassified. We will introduce a fourth distance r_b, which is the maximum allowable distance from the first root vector. We have used r_b to exclude vectors that corresponded to reflection from water.

As we calculate each row of the correlation matrix, we update the closest root vector. Let, g_j be defined by: $g_j = \max_k c_{jk}$. As each correlation coefficient is calculated, we can update g_j. Let $\varepsilon = 1 - r$. Then, $\varepsilon_a = 1 - r_a$. After we calculate each row of the correlation matrix, we look for a vector that is "not close" to any of the current root vectors and add it to the list of root vectors. Thus, the next root vector satisfies: $g_j < \varepsilon_a$. and $c_{1k} > \varepsilon_b$. When we get to the end of the list of vectors, we stop clustering.

We will store the root vectors in an array: $\phi(m)$. Before we begin clustering, we choose the first root vector [$\phi(1)$] and set the cluster counter (mc) to zero. During the cluster algorithm, we will choose the next root vector.

RESULTS

We have analyzed the hyperspectral images for the nine sites listed in Table 1. To compare Site B with its neighborhood, we expanded the area from 12,740 pixels to 90,720 pixels. The number of pixels ranges from 23,100 for Site A to 104,160 for Bluestem 134. The two rows for each of the sites present the results for two different cluster radii: 0.010 and 0.004. For the larger radius, the number of clusters ranges from 35 for Zink Ranch 127 to 133 for Bluestem 134. For the smaller radius, the number of clusters ranges from 100 for Site A to 418 for Bluestem 134.

Name	Clusters	Members	Associated	Unclustered	Pixels
Blue 134	133	76120	28027	13	104160
	418	55271	48785	104	104160
Blue 135	81	53687	21366	19	75072
	238	31379	43625	68	75072
Blue 136	61	49244	10755	1	60000
	185	27522	32454	24	60000
Site A	46	13702	3337	6061	23100
	100	7573	8856	6671	23100
Site B	82	68290	11202	11288	90720
	219	44506	34709	11505	90720
Old Site B	38	8787	2427	1526	12740
TGP 118	49	20324	8053	9	28416
	125	11737	16652	27	28416
TGP 122	74	25059	23951	10	49020
	257	25420	23566	34	49020
Zink 127	35	25923	11030	7	36960
	115	25821	11132	7	36960
Zink 131	50	68171	19724	9	87904
	174	41595	46298	11	87904

Table 1. Summary of the clustering results.

At the end of the clustering algorithm, we eliminate overlap by assigning each pixel to the closest root vector. We associate pixels that are not members of a cluster with the nearest root vector, if they are within the group radius (r_g) of a root vector. In Table 1, the Members column is the sum of all of the cluster members, the Associated column is the sum of all of the associated pixels, and the Unclustered column is the sum of all pixels that are neither members nor associates.

The Unclustered values are high for Sites A and B because we used r_b to exclude the water pixels.

USGS Site B

The USGS Site B is in the upper middle of Figure 2. The boundaries of old (12740 pixel) analysis region are displayed in Figure 3. Figure 3 is based on the same three band HyVista reflectance data as in Figure 2. However, the magnitude of the colors for each pixel has been increased until one of the colors is at the maximum value. We call it a "Bright Pixel" image. The unvegetated areas disturbed by oil production and changes in the lake level are orange on the image. As displayed in Table 1, we have created clusters. The next step is to produce geobotanical maps in which each cluster has a different color and the cluster position is plotted on a map that covers the same region as Figure 3. In Table 1, the number of clusters ranges from 35 to 418. We have created a 27 value color table by allowing each of the three primary colors (red, green, blue) have three values (zero, middle, high). Since it is difficult to distinguish all of the 27 colors on a computer screen and even harder on printed output, we cannot create a unique color for each of the (35 to 418) clusters. The old Site B region has 38 clusters. We can display all of the big clusters with a unique color if we set the definition of big high enough. The big clusters for Site B are displayed in Figure 4. For Site B, big means members greater than 4. For Bluestem 134 with 133 clusters, big means members greater than 80. We are primarily interested in brine scars. We can create an image where most vegetation is green, and we can use the colors for all of the brown (not green)

clusters. For a large number of brown clusters the color scale repeats after all of the colors are used. Figure 5 displays all of the brown clusters for Site B. One of our challenges is to differentiate between different sources of environmental damage. For this site, we want to distinguish roads from areas damaged by produced water spills. An enlargement of the road in the upper left area of the study area is displayed in Figure 6 and an enlargement of the area down stream from the produced water holding pit is shown in Figure 7. The holding pit is the white region in Figure 7 (and all of the other figures). From the pit, the site has a gentle slope to the lakeshore on the right. Comparing the two images, we find two clusters that are only in the road image, nine clusters that are only in the pit image, and 7 clusters that are in both images. Thus, we can distinguish road damage from produced water damage.

Site B. RGB. Bright pixel.

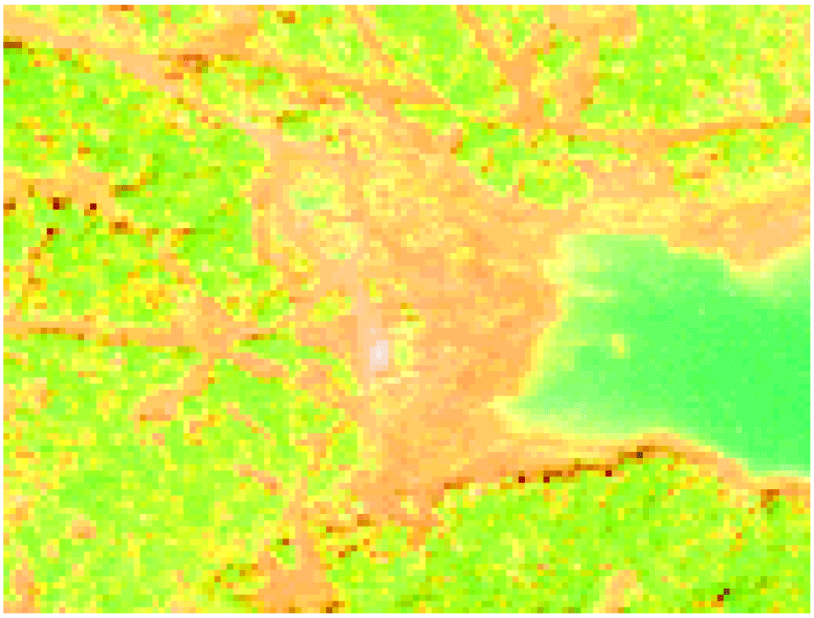

Figure 3. Bright Pixel Image of the Initial Site B Analysis Region.

Site B. Big Clusters (members > 4)

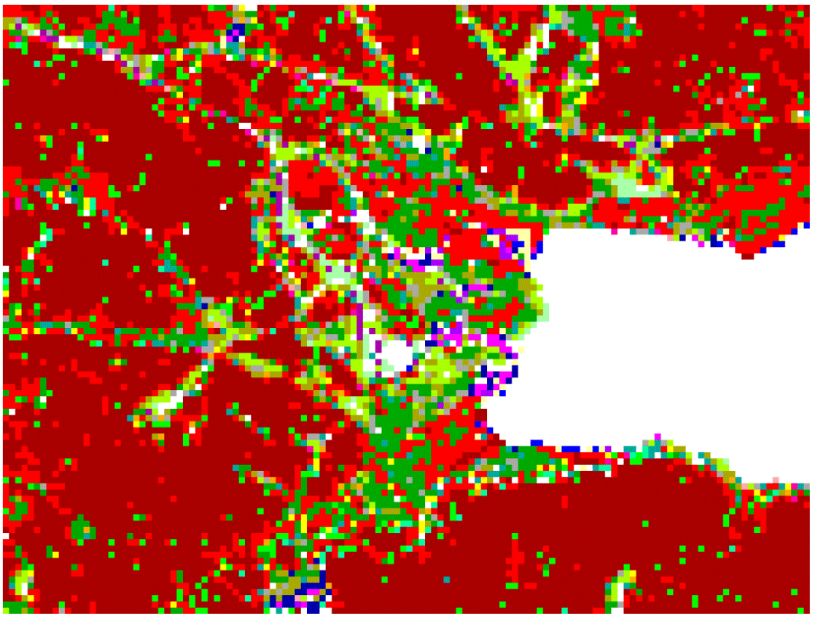

Figure 4. Site B Big Clusters.

Site B. All Clusters (members > 2)

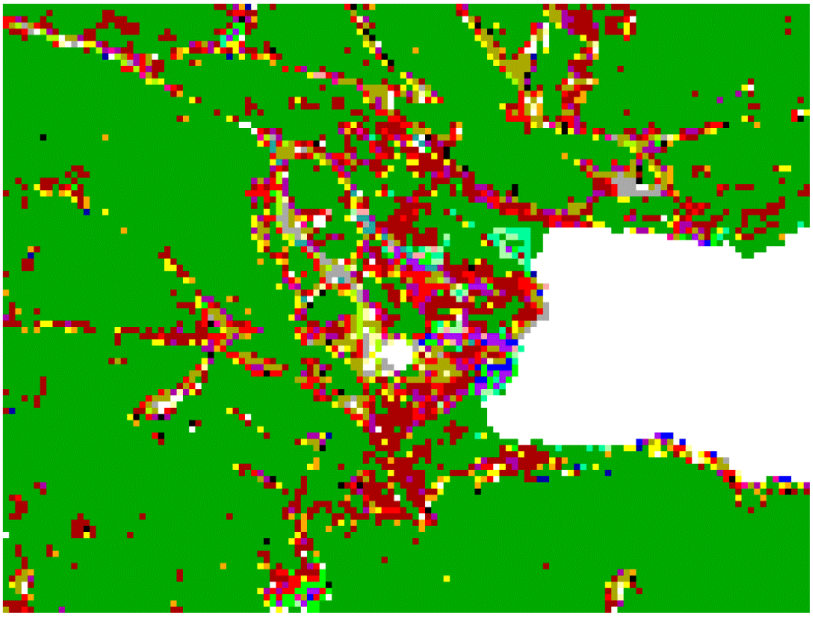

Figure 5. All Brown Clusters for Site B.

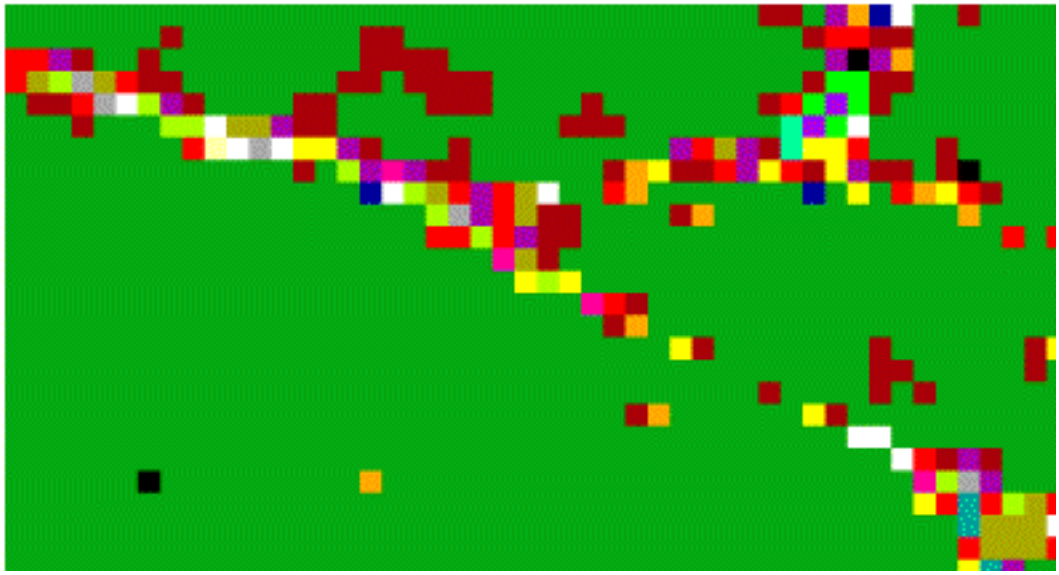

Figure 6. Enlargement of Road in upper left of Figure 5.

Figure 7. Enlargement for Site B.