REGION BASED DATA MINING ON AGRICULTURE DATA

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Babitha Battu

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

November 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Region Based Data Mining on Agriculture Data

By

Babitha Battu

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Anne Denton
Chair

Dr. Simone Ludwig

Dr. David Franzen

Approved:

11/12/2015	Dr. Brian M. Slator
Date	Department Chair

ABSTRACT

Spatial Data Mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial databases. Most relationships in spatial datasets are regional and there is a great need for regional regression methods that derive regional reflects different spatial characteristics of different regions. A central challenge in spatial data mining is the efficiency of spatial data mining algorithms, due to the often huge amount of spatial data and the complexity of spatial data types and spatial accessing methods. This paper proposes a regional regression technique for regions that are defined by a categorical attribute, in particular soil type. The result is a series of hierarchically grouped regions according to their similarity.

ACKNOWLEDGEMENTS

I would like, first and foremost, to thank my family for their valuable support and constant encouragement.

I would like to take this opportunity to thank my advisor, Dr. Anne Denton, who has given me valuable support, encouragement and advice without which this work would not have been completed. I am thankful to the members of the committee, Dr. Simone Ludwig and Dr. David Franzen, for their time and suggestions.

I would also like to thank my friends, for their support in successful completion of my paper.

This material is based upon work supported by the National Science Foundation through grants PFI-1114363 and IIA-1355466. We thank American Crystal Sugar Company for working with us on the sugar beet portion of the analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

With the development of database technology and information collection methods, especially the extensive use of the satellite remote sensing technology, the quantity and complexity of data is increasing rapidly. There is great potential for discovering useful information and knowledge from these large and complex databases, and spatial data mining is playing an increasingly important role in finding relations within these large datasets.
Spatial data mining is used to extract implicit knowledge, spatial relationships, as well as other non-explicit information from spatial database; - however the complexity of spatial data results in many challenges [1].

Agriculture is a major industry of North Dakota, with 90 percent of the state area is comprised of farmland. The economy of the state is highly dependent on agriculture [2]. North Dakota has diverse crops: several categories of wheat, soybeans, corn and barley, sunflower, canola, sugar beet, potato, flax, field pea, lentils, chickpea and several other minor crops, making it the most diverse state in terms of crops grown in the Great Plains of the USA [3]. Thirty-eight percent of the total economic base is shared by the agricultural sector [4]. Globally, the increase in population and rapid urbanization has led to decreasing farmlands. As a result, reduced agricultural area is charged with feeding more people. To keep up with the nutritional demands of the rising populations and in some cases increased income growth, global food production must increase by 70 percent from 2014 levels in order to be able to feed the world. One of the key factors to achieve this is the maximum utilization of farmland.

Today the concept and technique of farming has changed significantly. Today’s agriculture is based more on technology than labor. After the introduction of precision agriculture the total scenario of agriculture has changed. In 1990, Gambardella and Carlen defined precision agriculture as: by collecting real-time data on weather, soil and air quality, crop maturity and even equipment and labor costs and availability, predictive analytics can be used to make smarter decisions [5]. With precision agriculture, control centers collect and process data in real time to help farmers make the best decision with regard to planting, fertilizing and harvesting crops. Sensors placed throughout the fields are used to measure temperature and humidity of the soil and surrounding air. In addition, picture of fields are taken using satellite imagery and robotic drones. The images over time show crop maturity and when coupled with predictive weather modeling showing pinpoint conditions 48 hours in advance [5].

Region based data mining is a type of spatial data mining, where different regions and the relationships among those regions are identified. In this paper, an algorithm is developed to create a hierarchical structure showing relationships among the categorical attribute, soil type of an agriculture dataset. The hierarchical structure is created using the regression results of the soil type attribute.

1.1. Problem Statement

Region-based data mining is a type of spatial data mining, where different regions and the relationships among those regions are identified. In this paper, an algorithm is developed to identify hierarchical relationships in agriculture data among regions that are defined based on the categorical attribute soil type. The hierarchical structure is derived using the regression results of NDVI dependency on yield for a dataset in which regions differ by soil type. The result is shown as a dendrogram of clusters showing soil types in a hierarchical structure.

CHAPTER 2. LITERATURE SURVEY

Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. There is a need for effective and efficient methods to extract unknown information from spatial data sets of large size and complexity [6]. Due to the widespread application of geographic information systems (GIS) and GPS technology, private industries and the general public also have more and more interest in both contributing and using geographic data. Spatial data mining is still at a very early development stage and its limits and potentials are yet to be defined. In spatial data mining, the data cannot tell stories unless we formulate appropriate questions to ask and use appropriate methods to seek the answers from the data.

Pei et al. focuses on the development of a new method for point pattern analysis for detecting feature from spatial point processes using collective nearest neighbor [7]. Establishing spatial clustering methods are often sensitive to the parameterization of the clustering algorithm, particularly to the scale at which one theorizes clustering occurs, as such an assumption often must be made a priori to the application of the clustering technique. Consequently, the results of clustering may be highly subjective. To address this issue, Pei et al. present a new method of clustering they call the collective nearest neighbor (CLNN) method. The basis for CLNN is the distinction between points whose distribution may be explained by a causal mechanism versus those whose distribution may be explained by random ‘noise’, where the distinguishing characteristics between the two processes is intensity of clustering. CLNN extends previous research by developing a procedure for iterating over various scales of measurement to assess intensity. The authors demonstrate CLNN using both synthetic data as well as a case study focusing on identifying clusters of earthquakes in China from seismic data.

Geovisualization concerns the development of theory and method to facilitate knowledge construction through visual exploration and analysis of geospatial data and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication, and use [8]. As an emerging domain, geovisualization has drawn interests from various cognate fields and evolved along a diverse set of research directions, as seen in a recently edited volume on geovisualization by [9]. The main difference between traditional cartography and geovisualization is that, the former focuses on the design and use of maps for information communication and public consumption while the later emphasizes the development of highly interactive maps and associated tools for data exploration, hypothesis generation and knowledge construction [10].

In many application domains, data is collected and referenced by its geo-spatial location. Spatial data mining, or the discovery of interesting pattern in such databases, is an important capability in the development of database systems [11]. Presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypothesis to the end of better problem-solving and gaining deeper domain knowledge. The authors explain the importance of visual data mining on geo spatial data in a three step process: Overview first, zoom and filter, and then details on demand. Visualization technology is essential for presenting overviews and selecting interesting subsets. The visualizations of the data allow the data analyst to gain insight into the data, and thereby develop and confirm new hypotheses.

Some of the key advantages of visual data exploration over automatic data mining techniques alone are: it can deal with highly non-homogenous and noisy data, it can provide a qualitative overview of the data, it is useful when little is known about the data and the goals are vague. It yields results more quickly, with a higher degree of user satisfaction and confidence in the findings.

Spatial data mining describes objects or phenomena with specific real-world locations. Spatial data mining methods can be useful to understand the spatial phenomena and to discover relationships between spatial and non-spatial data. A common approach to analyzing geo-spatial data has been to apply standard statistical analysis methods. A significant problem in applying statistical methods to spatial data is that the models often assume or require statistical independence within the spatially distributed data. The difficulty is that spatial data items are often interrelated - objects are influenced by other, nearby objects. Regression models are applied to overcome this problem, but the overall analysis process is complicated.

Oner Ulvi Celepcikay et al. proposed a local statistical prediction model which recursively partitions data into small partitions and then fit a simple model to these small partitions [14]. The early Classification and Regression Tree (CART) algorithm [15] selects the split variable and split value that minimizes the weighted sum of the variances of the target values in the two subsets. The selection of first attribute to split in regression trees dramatically affects the resulted regions and that causes lack of flexibility. Since data is split greedily using a top-down approach, regions in regression trees are rectangular. Regression trees also aim to find local statistics, but our approach is more flexible since it employs an externally plugged-in fitness function to be maximized rather than evaluation variance of splitting on a single attribute like regression trees employ and also performs wider non-greedy search; moreover, shapes of regions that can be discovered by our approach can be convex polygons, which represent Voroinoi cells whereas regression trees are limited to discovery rectangle shape regions since they discover regions by recursively splitting trees into 2 sub-trees in a top down fashion.

Xun Zhou et al. has a naïve approach to sub-path discovery problem [16] and also proposes a Sub-path Enumeration and Pruning (SEP) approach with two design decisions on candidate sub-path traversal.

The Naïve approach has two phases, namely interesting sub-path identification, and dominated ISP elimination. In the first phase, the algorithm exhaustively enumerates sub-paths with all the length. It computes the distributive functions by scanning each sub-path entirely, and computing the interest measure. Then it determines the candidacy of ISP by computing the test. In the second phase, for each candidate ISP, the approach eliminates all the ISPs it dominates to generate the final result of DISPs and output the remaining ISPs.

The SEP approach addresses the two issues of Naïve approach:

1. The computation of the algebraic interest measure requires repetitive linear scans of each sub-path.
2. A large portion of the candidate sub-paths generated are actually dominated by other long sub-paths, which increases the time cost of the second phase.

For efficiently computing the aggregate functions, one solution is to materialize a lookup table of SUM function for sample data. Table 1 is an example look up table.

Table 1: Look table of SUM function in the sample data

Sub-path	**(1, 1)**	**(1, 2)**	**(1, 3)**	**(1, 4)**	**(1, 5)**	**(1, 6)**	**(1, 7)**	**(1, 8)**	**(1, 9)**	**(1, 10)**	**(1, 11)**
SUM	**7**	**1**	**2**	**1**	**6**	**11**	**15**	**12**	**17**	**22**	**12**

To reduce the computational cost, the goal is to achieve a constant-cost O(1) computation of aggregate function over any sub-path using the table, and limit the computational cost of building such a table to O(n). The second issue is addressed by partial-order traversal strategies.

CHAPTER 3. CONCEPTS

Data structures used for storing spatial data: Raster and vector are the two basic data structures for storing and manipulating images and graphics data on a computer.

3.1. GIS Data Models

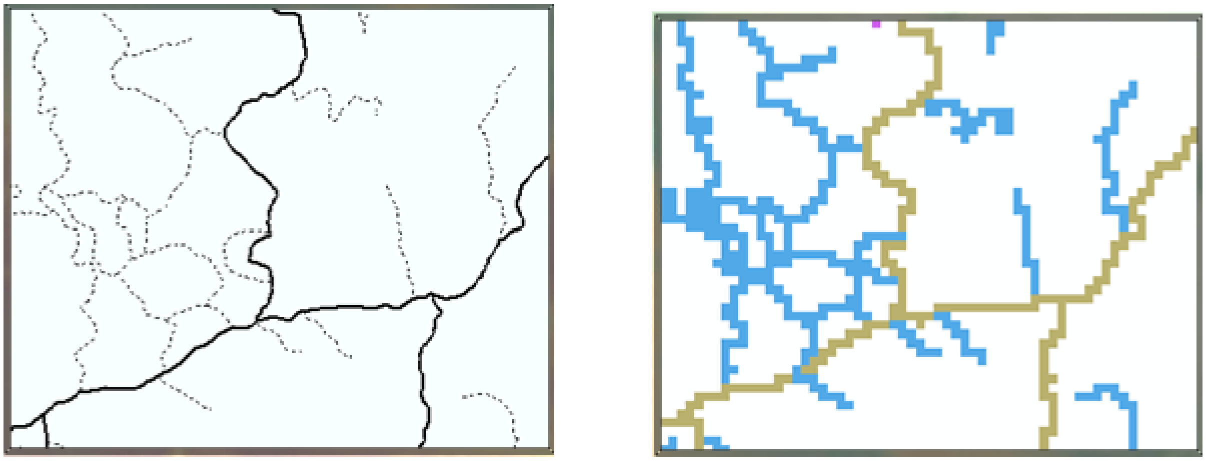


Figure 1: Vector Data Model and Raster Data Model

Raster images come in the form of individual pixels, and each spatial location or resolution element has a pixel associated where the pixel value indicates the attribute, such as color, elevation, or an ID number. Raster images are normally acquired by satellites, optical scanner, digital CCD camera and other raster imaging devices. Because a raster image has to have pixels for all spatial locations, it is strictly limited by how big a spatial area it can represent. When increasing the spatial resolution by 2 times, the total size of a two-dimensional raster image will increase by 4 times because the number of pixels is doubled in both X and Y dimensions [12]. The same is true when a larger area is to be covered when using same spatial resolution. Vector data can be easily converted to raster data. Figure 1, shows the graphical representation of raster and vector data models. Figure 2, shows the raster data model.

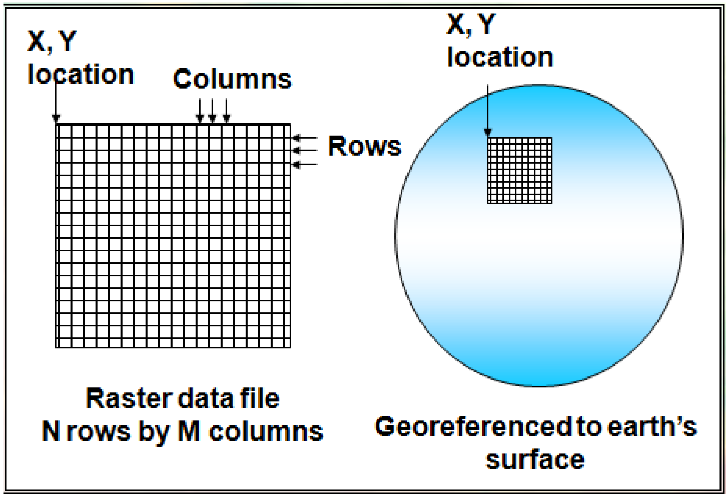


Figure 2: Raster Data Model

Vector data comes in the form of points, lines and polygons that are geometrically and mathematically associated [13]. Points are stored using the coordinates, for example, a two-dimensional point is stored as (x, y). Lines are stored as a series of point pairs, where each pair represents a straight line segment, for example, (x1, y1) and (x2, y2) indicating a line from (x1, y1) to (x2, y2). Figure 3, shows the vector data model in x and y planes.

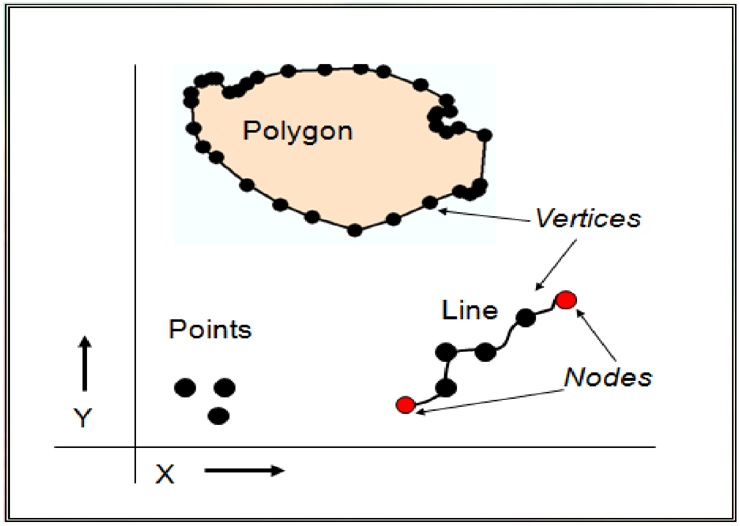


Figure 3: Vector Data Model

Vector data sets and raster data sets are both important in a GIS. Each has its strengths, therefore, it is counterproductive to use just one form of these datasets. An important difference between these two data sets is noticeable in the visualization of the data set.

3.2. Yield (Crop Yield)

Yield is defined as a measurement of the amount of a crop that was harvested per unit of land area. Crop yield is the measurement often used for cereal, grain or legume and is normally measured in metric tons per hectare (or kilograms per hectare). Crop can also refer to the actual seed generation from the plant. For example, grain of wheat yielding three new grains of wheat would have a crop yield of 1:3 [14].

To estimate the crop yield, producers usually count the amount of a given crop harvested in a sample area. The harvested crop is then weighed, and the crop yield of the entire field is extrapolated from the sample.

The yield within a given field is not always same or equal throughout the field. This may be due to soil variability, different amounts of fertilizers, pesticides and herbicides input requirement at different location within the field. If these problems are addressed site specifically we can expect average and better yield throughout the field. The first and basic study for this variability can be the classified yield map, a map showing yield variation within a field. A yield map has a great value in planning and management of crop production. The yield within a specific field is not always similar or average throughout. Yield differences may be due to soil variability, different amount of fertilizer, and pesticides input and efficacy at different locations within the field. If these problems are addressed site specifically we can expect average and better yield throughout the field.

3.3. NDVI (Normalized Difference Vegetation Index)

The NDVI is an index of plant greenness or photosynthetic activity, and is the most commonly used vegetation index [17]. Vegetation indices are based on the observation that different surfaces reflect different types of light differently. Vegetation that is dead reflects more red light and less near infrared light. Likewise, non-vegetated surfaces have a much more even reflectance across the light spectrum. NDVI is calculated on a per-pixel basis as the normalized difference between the red and the near infrared bands from an image.

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{RED})}{(\mathrm{NIR} + \mathrm{RED})},$$

NIR is the near infrared band value for a cell and RED is the red band value for the cell. The output of NDVI is a new image file/layer. Values of NDVI can range from -1.0 to +1.0. Higher NDVI values signify active vegetation and low NDVI values mean there is little difference between the red and NIR signals, implying very little photosynthetic activity. Example values: 0.629, 0.09, 0.27 etc.

3.4. Regression Analysis

Linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [18]. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other, but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$

is the dependent variable. The slope of the line is ***b***, and ***a*** is y intercept (the value of y when x = 0).

3.5. Least-Squares Regression

The most common method for fitting a regression line is the method of least-squares [19]. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. To view the fit of the model to the observed data, one may plot the computed regression line over the actual data points to evaluate the results [19].

3.6. Root Mean Square Error

The regression line predicts the average y value associated with a given x value, that is also necessary to get a measure of the spread of the y values around that average. To do this, root-mean square error (RMSE) [20] is used.

To construct the RMSE, determine residuals. Residuals are the difference between the actual values and the predicted values. It is denoted as $\hat{y}_i$ - y_i, where y_i is the observed values for the i^{th} observation and $\hat{y}_i$ is the predicted value. They can be positive or negative as the predicted value under or over estimates the actual value. Squaring the residuals, averaging the squares, and taking the square root gives us the RMSE. This RMSE can be used as a measure of the spread of the y values about the predicted y value.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}$$

Example values: 1.291764e-16, 2.892073e-16, 9.609790e-16.

3.7. Distance Matrix

Distance matrix is a matrix (two-dimensional array) containing the distances, taken pairwise, between the elements of a set [21]. If there are *N* elements, this matrix will have size *N*x*N*.

Properties of distance matrix:

- The entries on the main diagonal are all zero (that is, a hallow matrix), i.e., $x_{ii} = 0$ for all $1 \leq i \leq N$.
- The matrix is a symmetric matrix ($x_{ij} = x_{ji}$)
- All the off-diagonal entries are positive ($x_{ij} > 0$ if $i \neq j$)

Graphical view of a distance matrix: In this image, black cells denote a distance of 0 and white as maximal distance. Figure 4, is an example graphical view of a distance matrix.

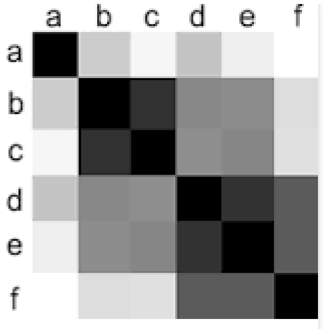


Figure 4: Graphical View of a Distance Matrix

3.8. Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of hierarchical clustering is:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item [22]. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeats steps 2 and 3 until all items are clustered into a single cluster of size N.

Figure 5, shows an example hierarchical dendrogram of six items.

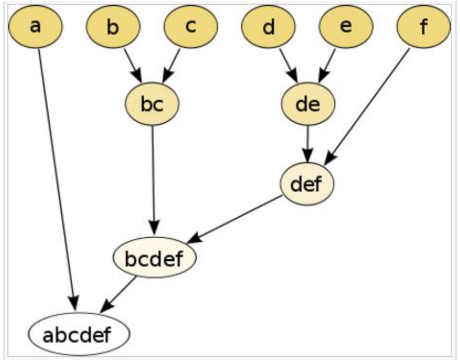


Figure 5: An Example Hierarchical Dendrogram Representation

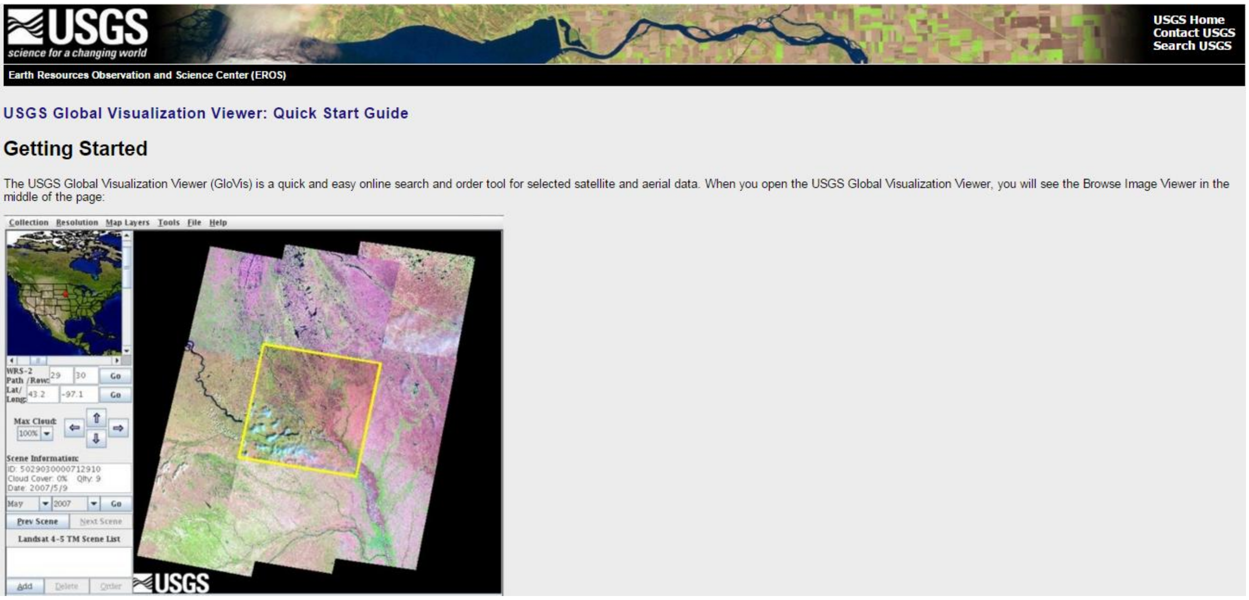
3.9. GIS

Over the past decade Geographical Information Systems (GIS) have evolved from a highly specialized niche to a technology that affects nearly every aspect of our lives, from finding driving directions to managing natural disasters.

3.10. GRASS GIS (Geographic Resources Analysis Support System)

GRASS GIS software suite is used for geospatial data management and analysis, image processing, graphics and maps production, spatial modeling and visualization. GRASS supports raster and vector data in two and three dimensions.

In this paper, we have used GRASS for calculating the NDVI for a raster map. This raster map should be in the form of .tiff (Tag Image File Format). From USGS (United Stated Geographical Survey) website [21], download the satellite (Landsat) image for the shape file under study (agricultural field). Figure 6, is the home page of the USGS site.



USGS
science for a changing world
USGS Home
Contact USGS
Search USGS
Earth Resources Observation and Science Center (EROS)

USGS Global Visualization Viewer: Quick Start Guide

Getting Started

The USGS Global Visualization Viewer (GloVis) is a quick and easy online search and order tool for selected satellite and aerial data. When you open the USGS Global Visualization Viewer, you will see the Browse Image Viewer in the middle of the page:

Collection Resolution Map Layers Tools File Help
Go
Go
Max Cloud:
100%
Go
Prev Scene
Next Scene
Landsat 4-5 TM Scene List
Add
Delete
Order
USGS

WRS-2 Path /Row: 29 30

Lat/ Long: 43.2 -97.1

Scene Information:
ID: 5029030000712910
Cloud Cover: 0% Qlty: 9
Date: 2007/5/9

May 2007

Figure 6: USGS Website to Download Landsat Images

For downloading an image, we require: longitude and latitudes of an image or the Path/Row of the required image. Select the Month and Year on which to capture the satellite image. The downloaded satellite image file names gives information regarding the type of the Landsat used. Table 2, shows example file name conventions used by different Landsat satellites.

Table 2: Landsat image file naming convention

File Name	Landsat
LM50310272012285EDC00	Landsat 5
LE70310272012229EDC00	Landsat 7
LC80310272013319LGN00	Landsat 8

Landsat represents the world’s longest continuously acquired collection of space-based moderate-resolution land remote sensing data [23].

Landsat 5 Thematic Mapper (TM) is recognized in the Guinness World Records for the longest operating earth observation satellite in history operating for 30 years, ceased in November 2011, acquisitions were initiated with the Multi-Spectral Scanner (MSS). Landsat 5 images consist of four spectral bands: blue, red, mid-infrared and near-infrared with a resolution of 30 meters.

Landsat 7’s Primary instrument is the Enhanced Thematic Mapper (ETM+). ETM+ added a panchromatic band with 15 m ground resolution (band 8). Landsat-7 continues to capture visible (reflected light) bands in the spectrum of blue, green, red, near-infrared (NIR) and mid-infrared (MIR) with 30 meter spatial resolution (bands 1-5, 7). It also has a thermal infrared channel with 60meter spatial resolution (band 6). Landsat 7 images resulted in partially missing data because of the SLC failure.

Landsat 8’s primary two sensors are the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). Generates 11 bands, seven out of eleven bands are consistent with Landsat7. Landsat 8 bands are: coastal, blue, green, red, NIR, SWIR-1, SWIR-2 and cirrus, with resolution of 30 meters [24]. The two new bands (band 10 & 11) are long wavelength infrared with resolution of 100 meters. Table 3, lists the band names along with their spatial resolution.

Table 3: Landsat spectral bands

Band	Name	Band width (λ, µm)	Spatial Resolution
1	Blue	0.45 - 0.515	30 m
2	Green	0.525 - 0.605	30 m
3	Red	0.63 - 0.69	30 m
4	Near Infrared	0.75 - 0.90	30 m
5	Shortwave IR-1	1.55 - 1.75	30 m
6	Thermal IR	10.4 - 12.5	60 m / 120 m*
7	Shortwave IR-2	2.09 - 2.35	30 m
8*	Panchromatic	0.52 - 0.9	15 m

3.11. ArcGIS

ArcGIS is used for working with maps and geographic information. It is used for: creating and using maps, compiling geographic data, analyzing mapped information, sharing and discovering geographic information and managing geographic information in a database.

3.12. ArcMap

ArcMap is the main component of Esri's ArcGIS suite of geospatial processing programs, and is used primarily to view, edit, create, and analyze geospatial data. ArcMap allows the user to explore data within a data set, symbolize features accordingly, and create maps.

3.13. R

R is a language and environment for statistical computing and graphics. R has variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible [25].

R has an effective data handling and storage facility. It has a large, coherent, integrated collection of intermediate tools for data analysis. Have graphical facilities for data analysis and is well developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

CHAPTER 4. IMPLEMENTATION

To identify regions in a field from agriculture data, we have used:

- Soil map: This map contains data about soils present in the field. It is in the form of a shape file (vector), collected from soil sensors and is collected from a farmer of Jamestown.
- Yield map: This map is in the form of a shape file (vector), with dry yield values in bushels per acre. This yield data is collected from the agriculture fields using yield monitors. The source of this file is from a farmer at Jamestown.
- Raster map: The source of this map is USGS (United States Geographical Survey). Download the Landsat 7 image for the corresponding yield map. This map is used to calculate the NDVI values [21].
- Field boundary map: This data is collected with the help of field sensors. Field boundary is used to set the extent of data frame.

The procedure followed to identify regions based on soil types is explained below:

4.1. Data Collection

Import the raster data in GRASS and calculate the NDVI values using GRASS. We have used a script to generate raster map with NDVI values.

Import all the required data to ArcGIS for further processing. Convert both yield and NDVI data to point data to extract the point value, these point maps are then spatially joined with each other.

From the given agriculture data, we could identify that all the soil types are loamy soils. Gardeners are advised that a loamy garden soil is best for just about all plants. Table 4, shows the abbreviated soil names for the soil types in the experimental agriculture data.

Table 4: Soil data used for experiment

Soil_Name	**Soil_Type**
BBDL39slope	Buse-Barnes-Darnen Loams, 3-9 slope
BBL36slope	Barnes-Buse Loams, 3-6 slope
BBLL69slope	Barnes-Buse-Langhei Loams, 6-9 slope
BSL36slope	Barnes-Svea Loams, 3-6 slope
BSL03slope	Barnes-Svea Loams, 0-3 slope
TSL01slope	Tonka Silt Loam, 0-1 slope
HWL03slope	Hamerly-Wyard Loams, 0-3 slope
SCL03slope	Svea-Cresbard Loams, 0-3 slope
FRL02slope	Fordville-Renshaw Loams, 02 slope
SARC69slope	Sioux-Arvilla-Renshaw complex, 6-9 slope
CFSL02slope	Clontarf Fine Sandy Loam, 0-2 slope
LPFCC02	La Prairie-Fluvaquents, Channeled Complex, 0-2
HTLFS06slope	Hecla-Towner Loamy Fine Sands, 0-6 slope
CFSL26slope	Clontarf Fine Sandy Loam, 2-6 slope

4.2. What Is Loamy Soil?

Soil is composed of many particles of varying sizes. Soil scientists have classified soil particles into three major groups: sand, silt and clay. Sand particles are the largest and tend to hold little water but allow good aeration. Clay particles are very small in size and tend to pack down so that water does not drain well and little or no air can penetrate. Silt particles are medium sized and have properties in between those of sand and clay [26].

A loam soil, - is one that combines all three of these types of particles in specific sand, silt and clay concentration based on the textural triangle below. A loam is one of several particularly well-suited textures for maximum crop production because it holds plenty of moisture but also drains well so that sufficient air can reach the roots.

Figure 7 is used to define soil texture [27]:

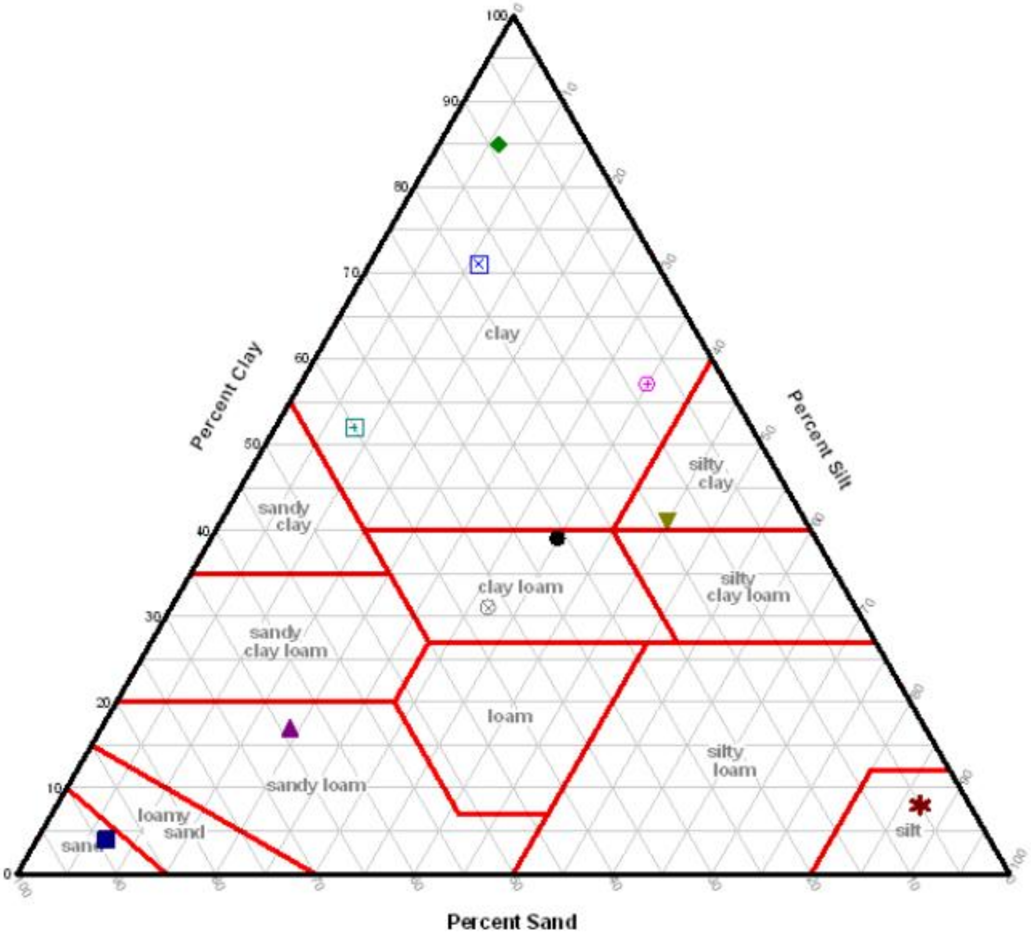


Figure 7: Soil Classification and Soil Moisture Estimation

4.3. Analysis of Map Data

The following figures shows the steps followed for implementation of algorithm:

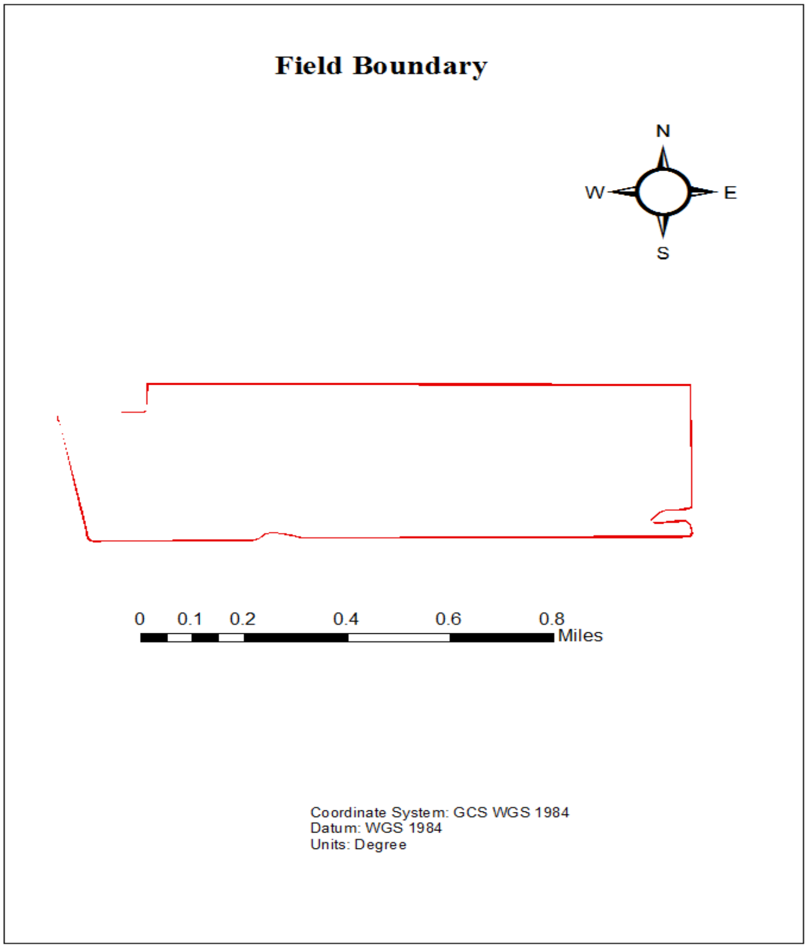


Figure 8: Field Boundary

Figure 8, is a map showing the field boundary of experimental agriculture data of a soybean field near Jamestown, North Dakota. This map defines the boundaries of the field. The data points which fall outside of this boundary are outliers and can be ignored for data analysis.

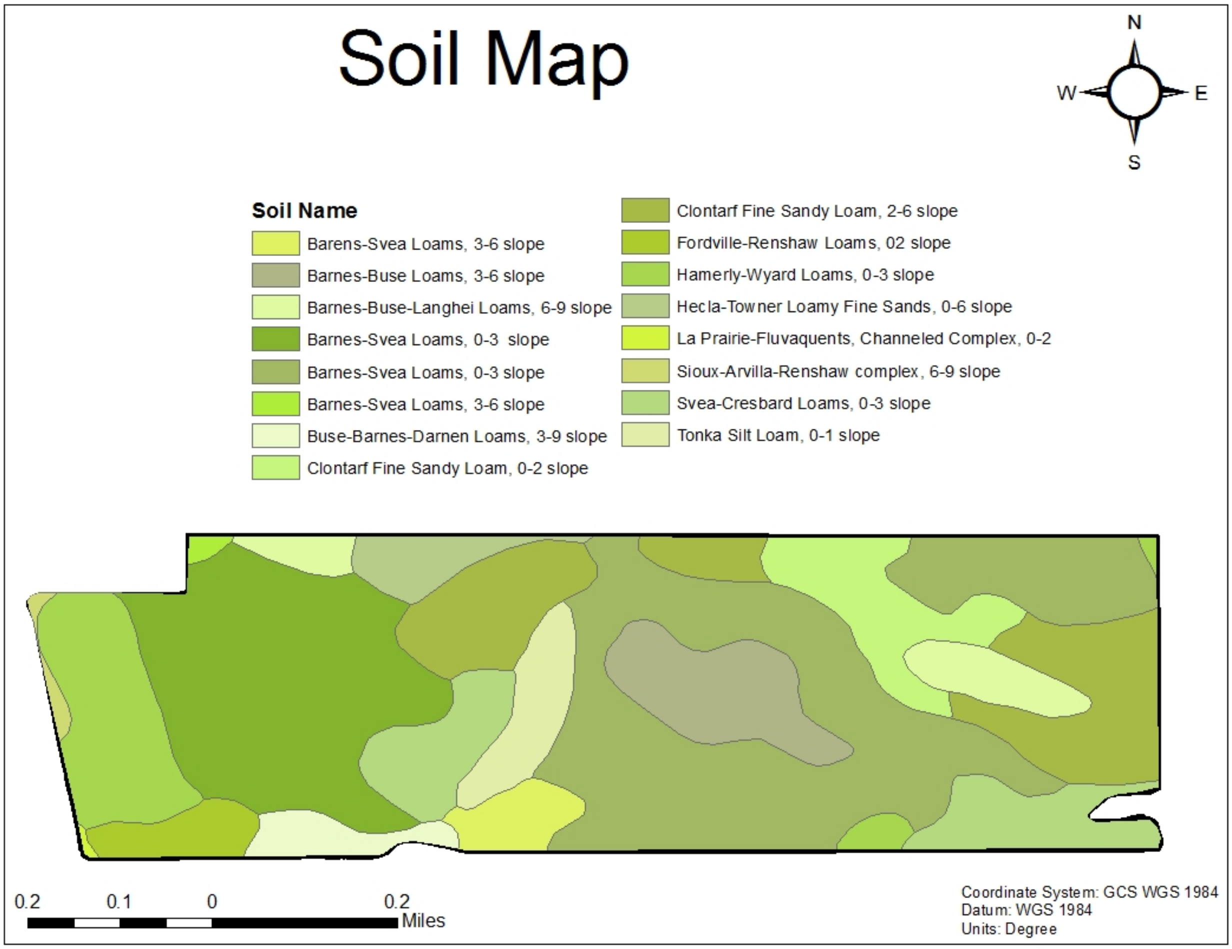

Soil Map
N
W
E
S
Soil Name
Barens-Svea Loams, 3-6 slope
Barnes-Buse Loams, 3-6 slope
Barnes-Buse-Langhei Loams, 6-9 slope
Barnes-Svea Loams, 0-3 slope
Barnes-Svea Loams, 0-3 slope
Barnes-Svea Loams, 3-6 slope
Buse-Barnes-Darnen Loams, 3-9 slope
Clontarf Fine Sandy Loam, 0-2 slope
Clontarf Fine Sandy Loam, 2-6 slope
Fordville-Renshaw Loams, 02 slope
Hamerly-Wyard Loams, 0-3 slope
Hecla-Towner Loamy Fine Sands, 0-6 slope
La Prairie-Fluvaquents, Channeled Complex, 0-2
Sioux-Arvilla-Renshaw complex, 6-9 slope
Svea-Cresbard Loams, 0-3 slope
Tonka Silt Loam, 0-1 slope
0.2
0.1
0
0.2
Miles
Coordinate System: GCS WGS 1984
Datum: WGS 1984
Units: Degree

Figure 9: Soil Zones with Soil Names

Figure 9, is the Soil data obtained in the form of polygon shape file showing variable soil types within the field. The obtained shape file was dissolved with the attribute soil name to aggregate them on basis of soil type. The dissolved shape was converted to raster with help of polygon to raster tool available in spatial analyst (ArcMap).

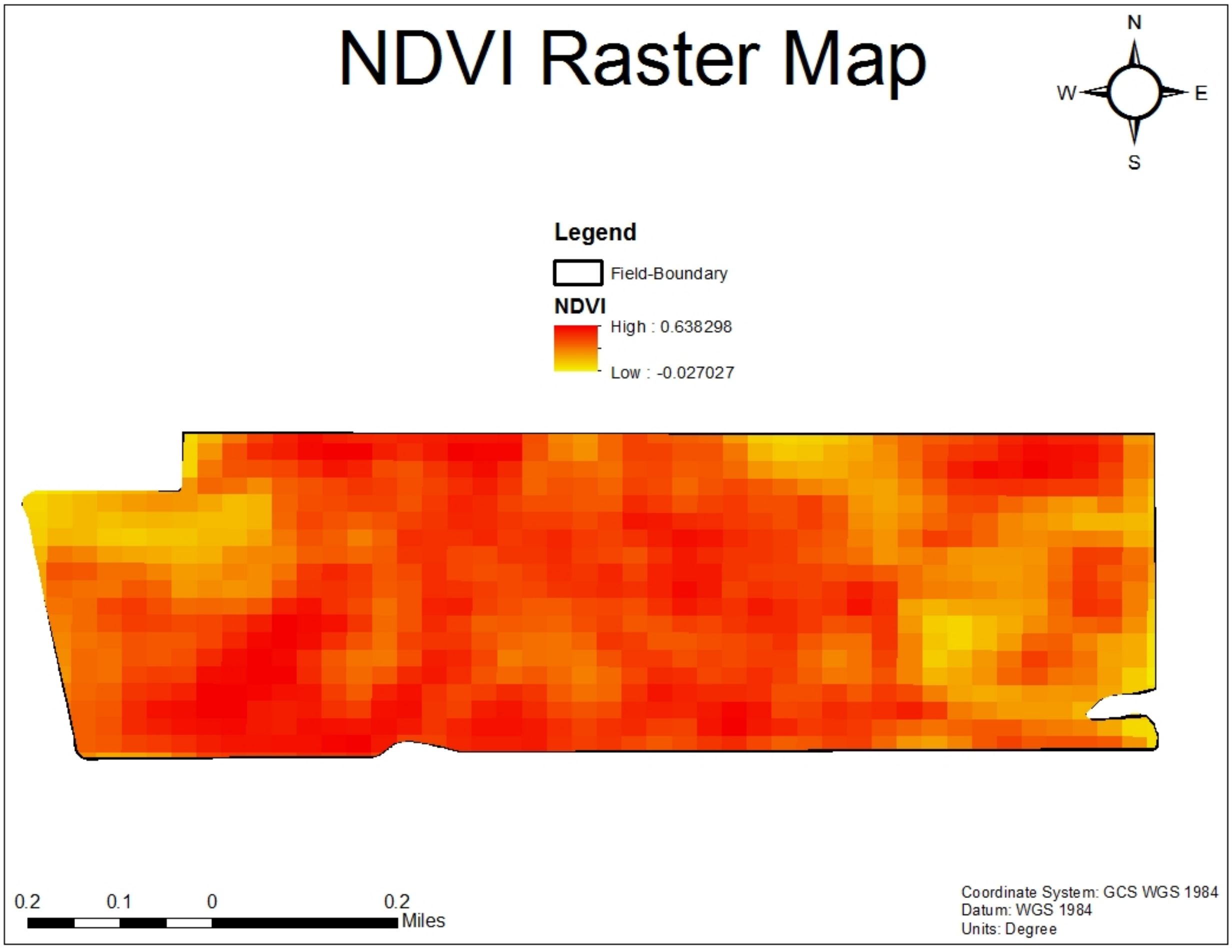


Figure 10: NDVI Map

Figure 10, is the calculated NDVI for the raster map (.tiff) downloaded from Landsat. The NDVI is calculated by running a script in GRASS. Then the map is imported in ArcMap for further analysis. And also the map is clipped to fit the field boundary by removing the pixels falling outside the field boundary.

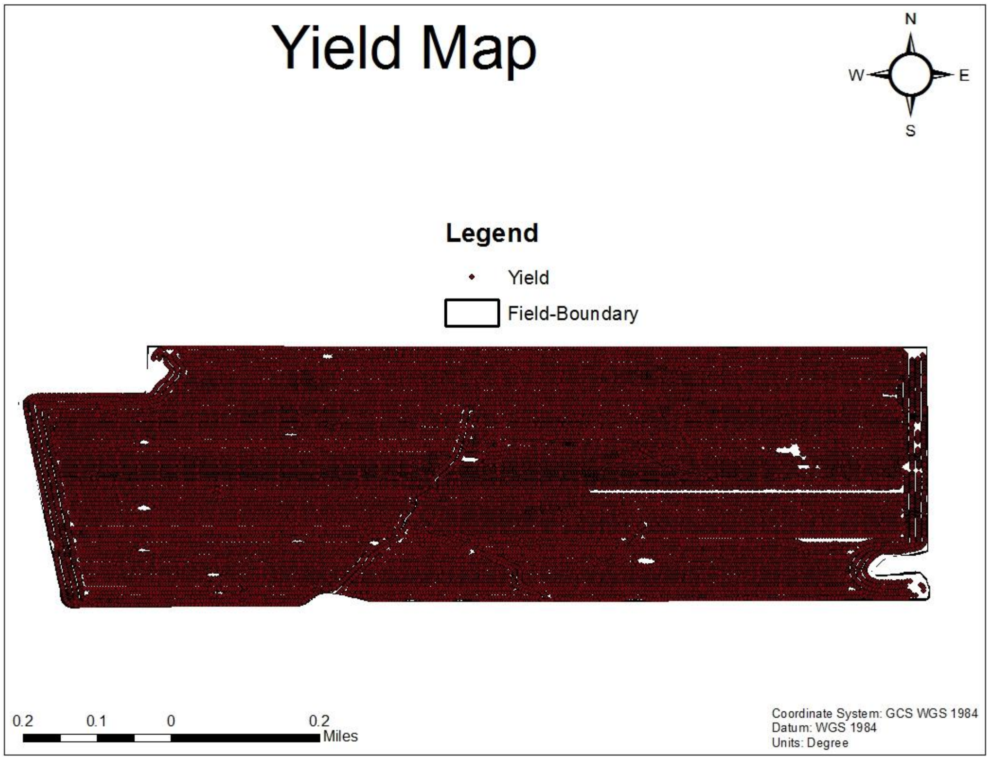


Figure 11: Yield Map

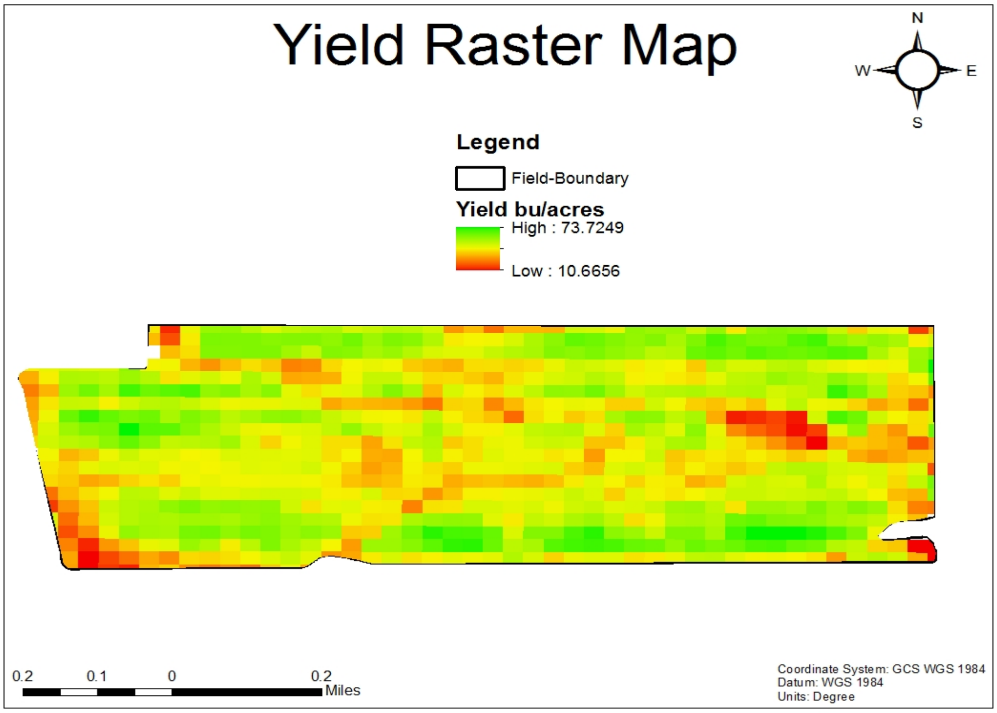


Figure 12: Yield Raster Map

The dry yield volume is in point shape file. The obtained values are in bushels per acre. One bushel per acre is equal to 0.06725 metric tons per hectare. This point shape file needs to be converted to raster for getting the yield value of each pixel. This is used for retrieving the corresponding NDVI value for the same pixel. Figure 12, is the map with yield map converted to raster.

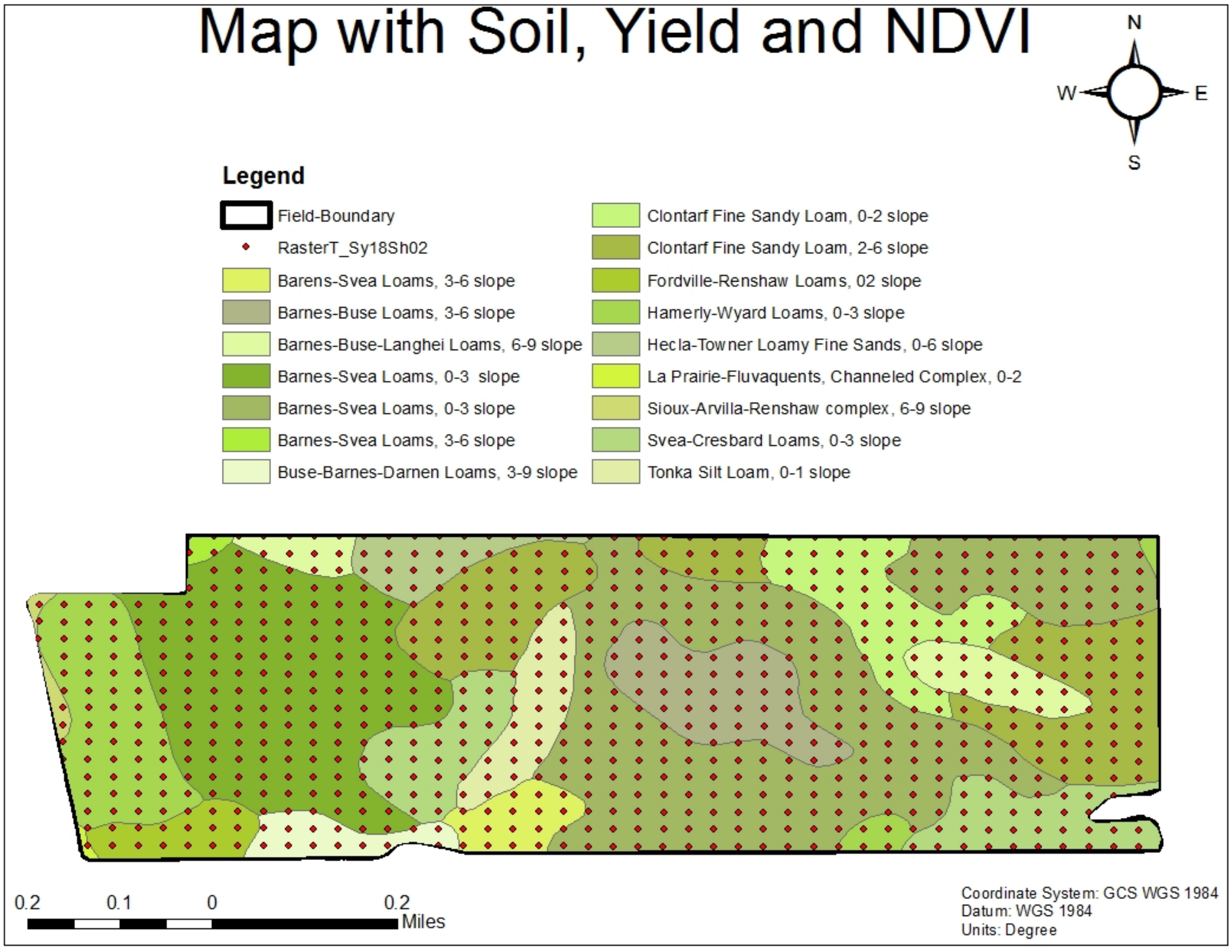

Map with Soil, Yield and NDVI
N
W
E
S
Legend
Field-Boundary
RasterT_Sy18Sh02
Barens-Svea Loams, 3-6 slope
Barnes-Buse Loams, 3-6 slope
Barnes-Buse-Langhei Loams, 6-9 slope
Barnes-Svea Loams, 0-3 slope
Barnes-Svea Loams, 0-3 slope
Barnes-Svea Loams, 3-6 slope
Buse-Barnes-Darnen Loams, 3-9 slope
Clontarf Fine Sandy Loam, 0-2 slope
Clontarf Fine Sandy Loam, 2-6 slope
Fordville-Renshaw Loams, 02 slope
Hamerly-Wyard Loams, 0-3 slope
Hecla-Towner Loamy Fine Sands, 0-6 slope
La Prairie-Fluvaquents, Channeled Complex, 0-2
Sioux-Arvilla-Renshaw complex, 6-9 slope
Svea-Cresbard Loams, 0-3 slope
Tonka Silt Loam, 0-1 slope
0.2
0.1
0
0.2
Miles
Coordinate System: GCS WGS 1984
Datum: WGS 1984
Units: Degree

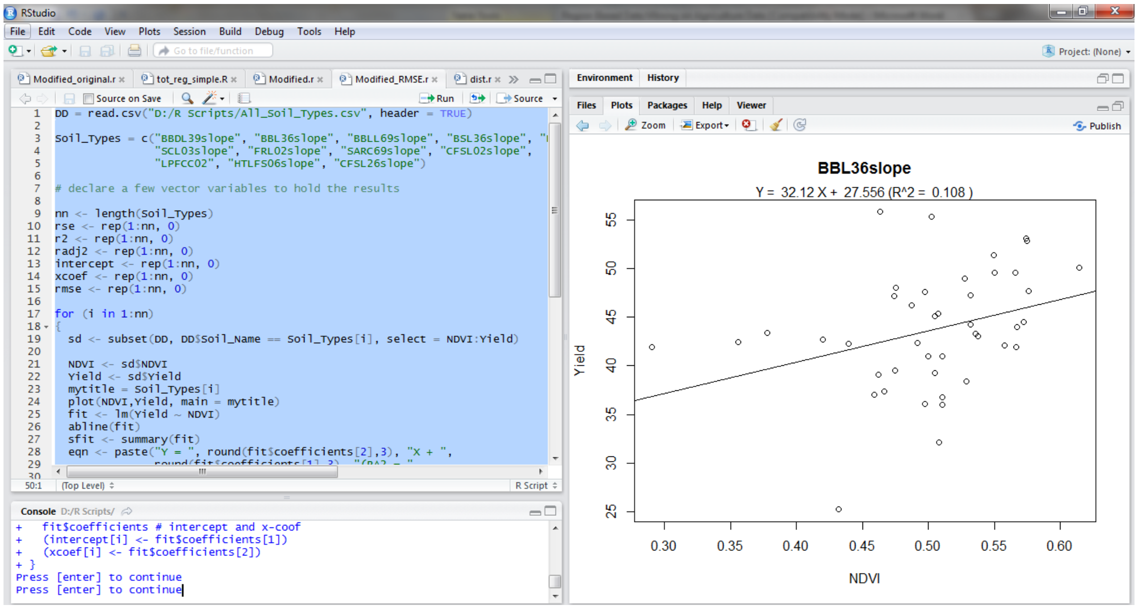
Figure 13: Analysis Map

After getting an image with yield and NDVI for soil types, copy the attribute table of this map and save it as a .csv (Comma Separate Values) file. For analysis on this data, it has to be loaded into the R Studio. We have written an R script to analyze the uploaded agriculture data and perform few statistical analyses creating plots for each soil type (region), calculate and draw regression lines for each plot, calculate the summary of the plot, and form regression line. Apart from the regression analysis, we have calculated Root Mean Square Error (RMSE) for each soil type and for the entire field. This RMSE is used as measure to identity relationships among the regions in a field.

The script created for performing the above statistics is:

```
DD = read.csv("D:/R Scripts/All_Soil_Types.csv", header = TRUE)
Soil_Types = c("BBDL39slope", "BBL36slope", "BBLL69slope", "BSL36slope",
               "BSL03slope", "TSL01slope", "HWL03slope",
               "SCL03slope", "FRL02slope", "SARC69slope", "CFSL02slope",
               "HTLFS06slope", "CFSL26slope")
# declare a few vector variables to hold the results
nn <- length(Soil_Types)
rse <- rep(1:nn, 0)
r2 <- rep(1:nn, 0)
radj2 <- rep(1:nn, 0)
intercept <- rep(1:nn, 0)
xcoef <- rep(1:nn, 0)
rmse <- rep(1:nn, 0)
for (i in 1:nn)
{
  sd <- subset(DD, DD$Soil_Name == Soil_Types[i], select = NDVI:Yield)
  NDVI <- sd$NDVI
  Yield <- sd$Yield
  mytitle = Soil_Types[i]
  plot(NDVI,Yield, main = mytitle)
  fit <- lm(Yield ~ NDVI)
  abline(fit)
  sfit <- summary(fit)
  eqn <- paste("Y = ", round(fit$coefficients[2],3), "X + ", round(fit$coefficients[1],3), "(R^2 = ",
            round(sfit$r.squared,3),")")
  xtextpos <- 1.4*min(NDVI)
  mtext(eqn, side = 3)
  #  text(xtextpos, min(Yield), labels = eqn)
  invisible(readline(prompt="Press [enter] to continue"))
  (r2[i] <- sfit$r.squared)
  (radj2[i] <- sfit$adj.r.squared)
  (rse[i] <- sfit$sigma)       # residual standard error
  (rmse[i] <- sqrt(mean(sfit$residuals)^2))
  fit$coefficients # intercept and x-coof
  (intercept[i] <- fit$coefficients[1])
  (xcoef[i] <- fit$coefficients[2])
}
```

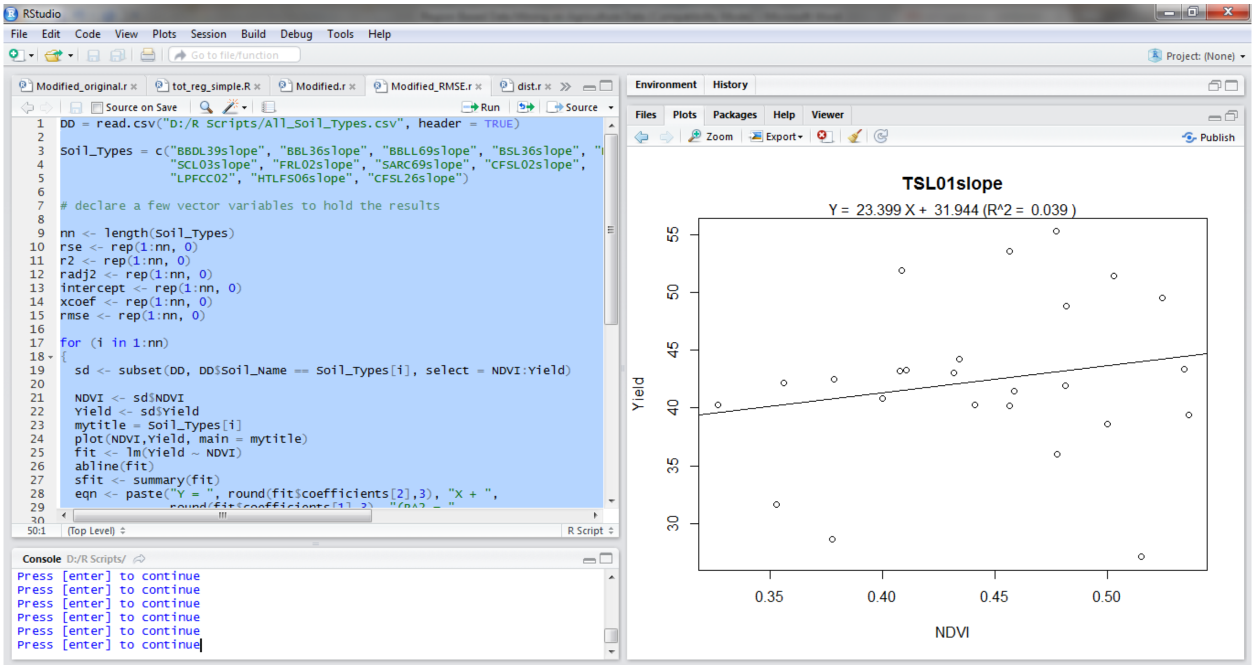
This script results in plots with regression lines along with the linear regression equation for each soil type (region) identified in the field.



```
DD = read.csv("D:/R Scripts/All_Soil_Types.csv", header = TRUE)

Soil_Types = c("BBDL39slope", "BBL36slope", "BBLL69slope", "BSL36slope", "
               "SCL03slope", "FRL02slope", "SARC69slope", "CFSL02slope",
               "LPFCC02", "HTLFS06slope", "CFSL26slope")

# declare a few vector variables to hold the results

nn <- length(Soil_Types)
rse <- rep(1:nn, 0)
r2 <- rep(1:nn, 0)
radj2 <- rep(1:nn, 0)
intercept <- rep(1:nn, 0)
xcoef <- rep(1:nn, 0)
rmse <- rep(1:nn, 0)

for (i in 1:nn)
{
  sd <- subset(DD, DD$Soil_Name == Soil_Types[i], select = NDVI:Yield)

  NDVI <- sd$NDVI
  Yield <- sd$Yield
  mytitle = Soil_Types[i]
  plot(NDVI,Yield, main = mytitle)
  fit <- lm(Yield ~ NDVI)
  abline(fit)
  sfit <- summary(fit)
  eqn <- paste("Y = ", round(fit$coefficients[2],3), "X + ",
               round(fit$coefficients[1],3), "(R^2 = "
```

```
+   fit$coefficients # intercept and x-coof
+   (intercept[i] <- fit$coefficients[1])
+   (xcoef[i] <- fit$coefficients[2])
+ }
Press [enter] to continue
Press [enter] to continue
```

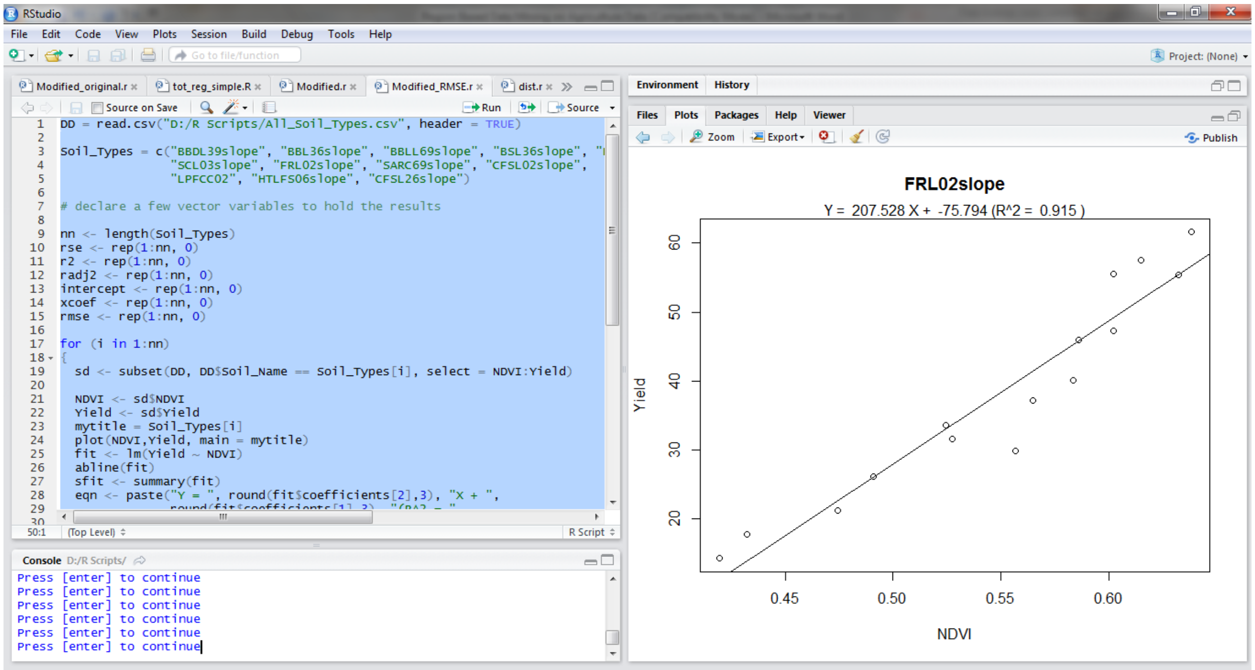
Figure 14: Plot for Soil Name: Barnes-Buse Loams, 3-6 slope



```
DD = read.csv("D:/R Scripts/All_Soil_Types.csv", header = TRUE)

Soil_Types = c("BBDL39slope", "BBL36slope", "BBLL69slope", "BSL36slope", "
               "SCL03slope", "FRL02slope", "SARC69slope", "CFSL02slope",
               "LPFCC02", "HTLFS06slope", "CFSL26slope")

# declare a few vector variables to hold the results

nn <- length(Soil_Types)
rse <- rep(1:nn, 0)
r2 <- rep(1:nn, 0)
radj2 <- rep(1:nn, 0)
intercept <- rep(1:nn, 0)
xcoef <- rep(1:nn, 0)
rmse <- rep(1:nn, 0)

for (i in 1:nn)
{
  sd <- subset(DD, DD$Soil_Name == Soil_Types[i], select = NDVI:Yield)

  NDVI <- sd$NDVI
  Yield <- sd$Yield
  mytitle = Soil_Types[i]
  plot(NDVI,Yield, main = mytitle)
  fit <- lm(Yield ~ NDVI)
  abline(fit)
  sfit <- summary(fit)
  eqn <- paste("Y = ", round(fit$coefficients[2],3), "X + ",
               round(fit$coefficients[1],3), "(R^2 = "
```

```
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
```

Figure 15: Plot for Soil Name: Tonka Silt Loam, 0-1 slope



```
DD = read.csv("D:/R Scripts/All_Soil_Types.csv", header = TRUE)

Soil_Types = c("BBDL39slope", "BBL36slope", "BBLL69slope", "BSL36slope", "
               "SCL03slope", "FRL02slope", "SARC69slope", "CFSL02slope",
               "LPFCC02", "HTLFS06slope", "CFSL26slope")

# declare a few vector variables to hold the results

nn <- length(Soil_Types)
rse <- rep(1:nn, 0)
r2 <- rep(1:nn, 0)
radj2 <- rep(1:nn, 0)
intercept <- rep(1:nn, 0)
xcoef <- rep(1:nn, 0)
rmse <- rep(1:nn, 0)

for (i in 1:nn)
{
  sd <- subset(DD, DD$Soil_Name == Soil_Types[i], select = NDVI:Yield)

  NDVI <- sd$NDVI
  Yield <- sd$Yield
  mytitle = Soil_Types[i]
  plot(NDVI,Yield, main = mytitle)
  fit <- lm(Yield ~ NDVI)
  abline(fit)
  sfit <- summary(fit)
  eqn <- paste("Y = ", round(fit$coefficients[2],3), "X + ",
               round(fit$coefficients[1],3), "(R^2 = "
```

```
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
Press [enter] to continue
```

Figure 16: Plot for Soil Name: Fordville-Renshaw Loams, 02 slope

The above figures show the plots for few soil types with their regression lines and their regression equations.

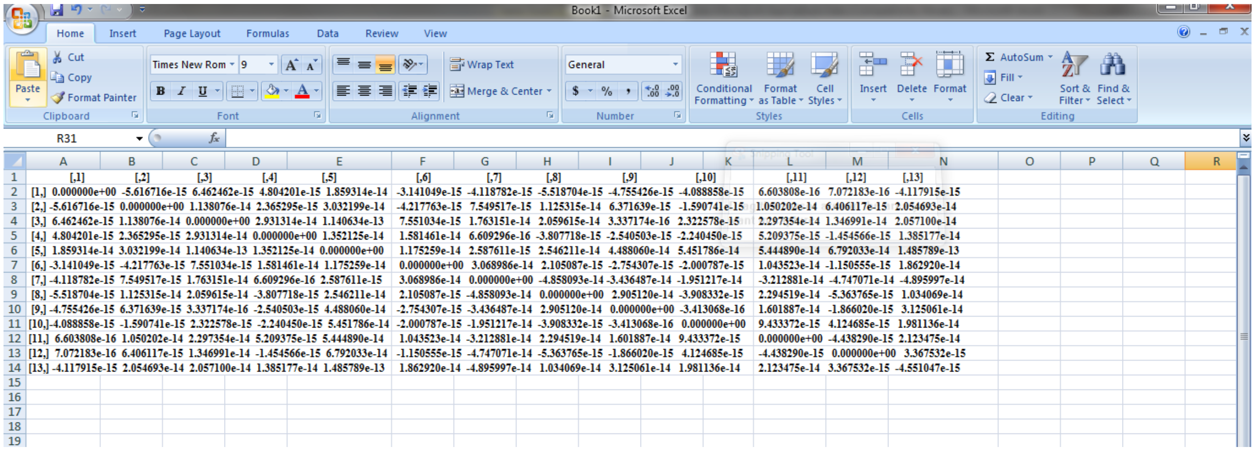
After calculating the RMSE values for each soil type, calculate distance matrix for all soil types using RMSE as a distance measure. The distance between any two regions is calculated with the formula:

$$\textbf{Distance } (S_i, S_j) = \textbf{RMSE } (S_i \text{ U } S_j) * (N_i + N_j) - (N_i * \textbf{RMSE } (S_i)) - (N_j * \textbf{RMSE } (S_j))$$

Where,

- S_i and S_j represent the two soil types for which the distance is calculated.
- N_i and N_j represent the number of rows or records (or values) in their respective regions.
- RMSE: Root Mean Square Error.

The resulting distance matrix must be a symmetric matrix with all diagonal values as 0's (zero).



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	0.000000e+00	-5.616716e-15	6.462462e-15	4.804201e-15	1.859314e-14	-3.141049e-15	-4.118782e-15	-5.518704e-15	-4.755426e-15	-4.088858e-15	6.603808e-16	7.072183e-16	-4.117915e-15
[2,]	-5.616716e-15	0.000000e+00	1.138076e-14	2.365295e-15	3.032199e-14	-4.217763e-15	7.549517e-15	1.125315e-14	6.371639e-15	-1.590741e-15	1.050202e-14	6.406117e-15	2.054693e-14
[3,]	6.462462e-15	1.138076e-14	0.000000e+00	2.931314e-14	1.140634e-13	7.551034e-15	1.763151e-14	2.059615e-14	3.337174e-16	2.322578e-15	2.297354e-14	1.346991e-14	2.057100e-14
[4,]	4.804201e-15	2.365295e-15	2.931314e-14	0.000000e+00	1.352125e-14	1.581461e-14	6.609296e-16	-3.807718e-15	-2.540503e-15	-2.240450e-15	5.209375e-15	-1.454566e-15	1.385177e-14
[5,]	1.859314e-14	3.032199e-14	1.140634e-13	1.352125e-14	0.000000e+00	1.175259e-14	2.587611e-15	2.546211e-14	4.488060e-14	5.451786e-14	5.444890e-14	6.792033e-14	1.485789e-13
[6,]	-3.141049e-15	-4.217763e-15	7.551034e-15	1.581461e-14	1.175259e-14	0.000000e+00	3.068986e-14	2.105087e-15	-2.754307e-15	-2.000787e-15	1.043523e-14	-1.150555e-15	1.862920e-14
[7,]	-4.118782e-15	7.549517e-15	1.763151e-14	6.609296e-16	2.587611e-15	3.068986e-14	0.000000e+00	-4.858093e-14	-3.436487e-14	-1.951217e-14	-3.212881e-14	-4.747071e-14	-4.895997e-14
[8,]	-5.518704e-15	1.125315e-14	2.059615e-14	-3.807718e-15	2.546211e-14	2.105087e-15	-4.858093e-14	0.000000e+00	2.905120e-14	-3.908332e-15	2.294519e-14	-5.363765e-15	1.034069e-14
[9,]	-4.755426e-15	6.371639e-15	3.337174e-16	-2.540503e-15	4.488060e-14	-2.754307e-15	-3.436487e-14	2.905120e-14	0.000000e+00	-3.413068e-16	1.601887e-14	-1.866020e-15	3.125061e-14
[10,]	-4.088858e-15	-1.590741e-15	2.322578e-15	-2.240450e-15	5.451786e-14	-2.000787e-15	-1.951217e-14	-3.908332e-15	-3.413068e-16	0.000000e+00	9.433372e-15	4.124685e-15	1.981136e-14
[11,]	6.603808e-16	1.050202e-14	2.297354e-14	5.209375e-15	5.444890e-14	1.043523e-14	-3.212881e-14	2.294519e-14	1.601887e-14	9.433372e-15	0.000000e+00	-4.438290e-15	2.123475e-14
[12,]	7.072183e-16	6.406117e-15	1.346991e-14	-1.454566e-15	6.792033e-14	-1.150555e-15	-4.747071e-14	-5.363765e-15	-1.866020e-15	4.124685e-15	-4.438290e-15	0.000000e+00	3.367532e-15
[13,]	-4.117915e-15	2.054693e-14	2.057100e-14	1.385177e-14	1.485789e-13	1.862920e-14	-4.895997e-14	1.034069e-14	3.125061e-14	1.981136e-14	2.123475e-14	3.367532e-15	-4.551047e-15

Figure 17: Distance Matrix for the Sample Agriculture Data

Once you have calculated the distance matrix, perform hierarchical clustering on the resultant distance matrix and create a dendrogram showing the relationship among various soil types. The script for the creation of distance matrix, performing hierarchical clustering and creating a dendrogram:

```
nn <- length(Soil_Types)
rmse_i <- rep(0, nn)
rmse_j <- rep(0, nn)
# Matrix to hold RMSE values
x = matrix(0, nn, nn)
for (i in 1:nn)
{
  sdi<- subset(DD, DD$Soil_Name == Soil_Types[i], select = NDVI:Yield)
  # Number of rows in current soil type
  si_nows = dim(sdi)[1]
  NDVI_i <- sdi$NDVI
  Yield_i <- sdi$Yield
  # Fitting Linear Models for carrying out Regression
  fit <- lm(Yield_i ~ NDVI_i)
```

```
# Summary of Linear Model
sfit <- summary(fit)
# Root Mean Square Error for the current soil type
rmse_i[i] <- sqrt(mean(sfit$residuals)^2)
  for(j in (i+1):nn)
  {
    if( j!= (nn+1))
    {
      sdj <- subset(DD, DD$Soil_Name == Soil_Types[j], select = NDVI:Yield)
      sj_nrows = dim(sdj)[1]
      NDVI_j <- sdj$NDVI
      Yield_j <- sdj$Yield
      fit_j <- lm(Yield_j ~ NDVI_j)
      sfit_j <- summary(fit_j)
      rmse_j[j] <- sqrt(mean(sfit_j$residuals)^2)
      # Concatinating two soil types
      s_ij = rbind(sdi,sdj)
      # Total number of rows after Concatenation
      tot_rows_ij = dim(s_ij)[1]
      # NDVI & Yield for the combined soil types
      NDVI <- s_ij$NDVI
      Yield <- s_ij$Yield
      fit_tot <- lm(Yield ~ NDVI)
      sfit_tot <- summary(fit_tot)
      # Formula for calculating distance between two soil types
      dist_sij = (sqrt(mean(sfit_tot$residuals)^2)) * tot_rows_ij - (si_nows * rmse_i[i])
      - (sj_nrows * rmse_j[j])
```

```
        x[i, j] = dist_sij
        x[j, i] = dist_sij
      }
    }
}
# Calculating Distance matrix for x
B = dist(x)
# Printing the Distance Matrix
print(B)
# Performing Heirarchical Clustering Distance Matrix
hc <- hclust(B,method = "average")
# Plotting the Dendrogram
plot(hc, labels = Soil_Types)
```

The output of the above script results in a dendrogram of regression results based on soil type attribute.

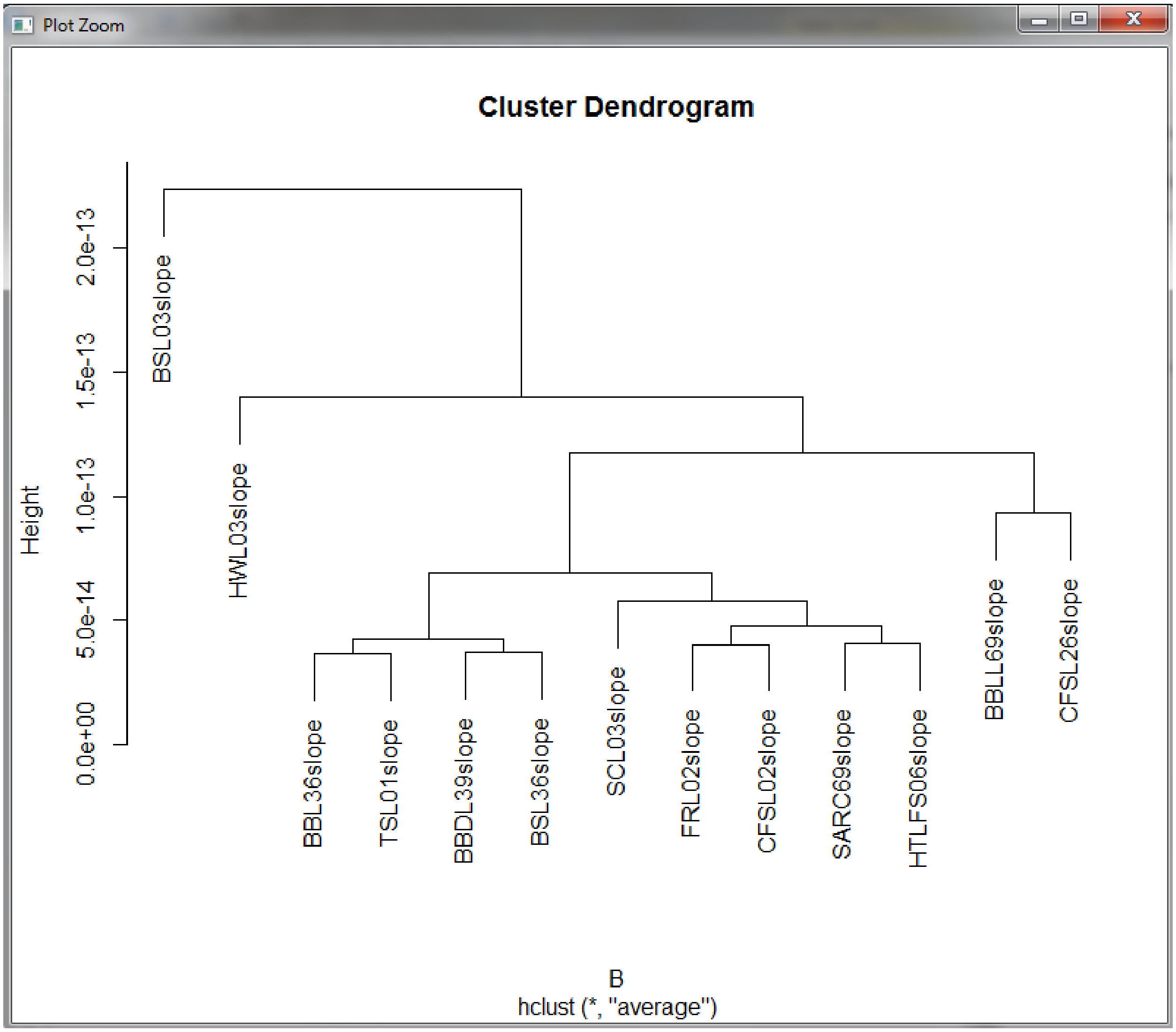


Figure 18: Hierarchical Clustering

The outcome of performing hierarchical clustering on the results of regression on soil type attribute shows, the regions which are close to each other are clustered together.

CHAPTER 5. CONCLUSIONS

As part of the work, we have shown four things. First, plots for each categorical attribute of agriculture data, i.e., soil type, showing regressions with regression line. Secondly, calculation of root mean square error for each soil type and constructing a distance matrix. Thirdly, the application of hierarchical clustering to produce the hierarchical structure. Finally, showing how soil types with similar NDVI relationship to grain yield are clustered close to each other. The experimental results show, - that soil types in a given cluster have similar component type values and their major components are similar. The algorithm used in this paper group data using a categorical attribute to cluster the regions based on the regression results than clustering based on NDVI and yield values. The algorithm successfully identified the hierarchical structure of regions within an agriculture data.

REFERENCES

1. Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." VLDB. Vol. 97. 1997

2. North Dakota Agriculture Stats. <http://www.farmflavor.com/us-ag/north-dakota/ >.

3. United States Department of Agriculture, National Agriculture Statistics Service. <http://www.nass.usda.gov>.

4. USGS for a changing world. <http://www.usgs.gov>.

5. Precision agriculture: Using predictive weather analytics to feed future generations. <http://www.research.ibm.com/articles/precision_agriculture.shtml>.

6. Mennis, Jeremy, and Diansheng Guo. "Spatial data mining and geographic knowledge discovery—An introduction." Computers, Environment and Urban Systems 33.6 (2009): 403-408.

7. Pei, T., Zhu, A. X., Zhou, C., Li, B., & Qin, C. Detecting feature from spatial point processes using collective nearest-neighbor. Computers, Environment and Urban Systems, 33(6), 2009: 435–447

8. MacEachren, A., & Kraak, M.-J. Research challenges in geovisualization. Cartography and Geographic Information Science, 2001: 283–312

9. Dykes, J., MacEachren, A. M., & Kraak, M.-J. Exploring geovisualization. Amsterdam: Elsevier, 2005

10. MacEachren, A. Visualization in modern cartography: Setting the agenda. In D. R. F. Taylor & A. M. MacEachren (Eds.), Visualization in modern cartography (pp. 1–12). Oxford, UK: Pergamon, 1994.

11. Keim, Daniel A., et al. "Pixel based visual data mining of geo-spatial data."Computers & Graphics 28.3, 2004: 327-344

12. GeoVITe – Geodata Visualization & Interactive Training Environment. <https://geodata.ethz.ch/resources/tutorials/L2GeodataStructuresAndDataModels/en/html/unit_u4VecVsRas.html>.

13. Raster Data and Vector Data. <http://www.highpointnc.gov/gis/raster_v_vector_data.cfm>

14. Crop Yield. <http://www.investopedia.com/terms/c/crop-yield.asp>.

15. Normalized Difference Vegetation Index. <https://en.wikipedia.org/wiki/Normalized_Difference_Vegetation_Index>.

16. REG^2: A Regional Regression Framework for Geo-Referenced Datasets. ACM GIS '09 November 4-6, 2009. Seattle, WA, USA (c) 2009

17. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees, Wadsworth, Belmont, CA, 1984

18. Zhou, Xun, et al. "Discovering interesting sub-paths in spatiotemporal datasets: A summary of results." Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, 2011

19. Least Squares Regression. <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>.

20. RMS Error. <http://statweb.stanford.edu/~susan/courses/s60/split/node60.html>.

21. Distance Matrix. <https://en.wikipedia.org/wiki/Distance_matrix>.

22. Hierarchical Clustering. <https://en.wikipedia.org/wiki/Hierarchical_clustering>.

23. Landsat Project Description. <http://landsat.usgs.gov//about_project_descriptions.php>.

24. Landsat Program: Satellite Imagery Data and Bands. <http://gisgeography.com/landsat-program-satellite-imagery-bands/>.

25. What is R?. <https://www.r-project.org/about.html>.

26. B. Rosie Lerner, What is Loam? <https://www.hort.purdue.edu/ext/loam.html>.

27. Soil Classification. <https://en.wikipedia.org/wiki/Soil_classification>.