SLIDING WINDOW BASED TECHNIQUE TO OBTAIN CORRELATION BETWEEN FIELD VARIABLES

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

By

Harshada Chandrakant Chavan

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

Major Department:
Computer Science

May 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Sliding Window Based Technique to Obtain Correlation between Field Variables

By

Harshada Chandrakant Chavan

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

John Nowatzki

Approved:

May 15, 2015
Date

Dr. Ken Magel
Department Chair

ABSTRACT

In agriculture, finding correlation between field variables such as yield, NDVI, etc. is a classic problem. The most popular solution for finding correlation is to use regression analysis over complete field. In a field, there are numerous soil variables such as soil composition, water content, etc. that affect the correlation and it is not always possible to consider such factors in regression model. These factors adversely affect the accuracy of the correlation model. We demonstrate that it is incomplete and inaccurate to represent such correlation using single regression model for the complete field. We propose a novel technique- Sliding Window Based Technique, which finds correlations over small areas, *windows*, of the field instead of modelling one correlation for entire field. We prove our claims with the help of experiments done on the field data. We evaluate the relationship over multiple window sizes and select appropriate window sizes for further analysis.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my adviser Dr. Anne Denton for introducing me to the field of geo-spatial data processing and her invaluable guidance throughout the course of my work. I would specially like to thank her for being patient to my basic questions and explaining things with great enthusiasm; it motivated me to explore this field even further. Working with her was a great learning experience.

I am thankful to John Nowatzki and Dr. David Franzen for always being available for technical discussions and giving me insight into the world of agriculture and soil science. I would like to take this opportunity to thank Johns group for providing yield and high resolution NDVI data.

I would like to acknowledge the support for this research by National Science Foundation and NDEPSCoR. This material is based upon work supported by the National Science Foundation through grants PFI-1114363 and IIA-1355466.

I am grateful to my committee members Dr. Simone Ludwig and John Nowatzki for their invaluable time.

I would like to mention my lab-mates and friends at North Dakota State University for making my stay so memorable.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Advancements in the techniques for mining data have extended the applicability of these techniques to problems from increasingly many domains. These techniques have been proved very effective at solving Big Data problems. Recent developments in agriculture are able to provide field data at frequent time intervals and with high precision. Such data is available in the form of, but not limited to, satellite images, output from various sensors or manually recorded information. The promise of satellite imagery has tremendously increased in the last decade because of at least two reasons [5]. Firstly, in 2008, United State Geological Survey decided to make the entire archive of Landsat data available [16] at no charge. Satellites recording these images are equipped with advanced preprocessing algorithms to geographically register the image which significantly reduces the time required to obtain images at relatively fine spatial resolution (30m X 30m). Secondly, numerous commercial satellites are offering even finer resolution (5m X 5m) at costs that are approaching 1 USD per km^2. In near future, high resolution data is going to be even more easily accessible, as large numbers of small satellites are being deployed [6]. An increased resolution of images results in a potential for an increase in the precision of data mining outcomes, but also an increase in complexity. Challenges for data mining in the agricultural domain are, therefore, expected to grow. In the recent past, scientists have already initiated the efforts to apply data mining techniques to agriculture [8].

In this work, we discuss drawbacks of the traditional use of linear regression and propose a novel way of using regression analysis to find correlation between multiple variables. We consider data from agricultural fields, in particular attributes such as yield, Normalized Difference Vegetation Index (NDVI), etc. We consider scenarios of simple linear regression only, however, the technique can be applied to any kind of regression analyses.

1.1. Motivation

Statistical methods such as regression techniques are commonly used for finding patterns in sets of data [17]. The purpose could be to study the change in the dependent variable as the independent variable changes or to predict the dependent variable for a specific instance of independent variable. An example of the need of finding such correlations is to find a relationship between yield and NDVI. From a farmer's point of view, it is important to identify regions on the field which are not producing high yield in spite of having high bio-mass (NDVI). Such analysis could help farmers to take corrective measures to improve the yield of these regions for example, applying more fertilizers. Also, the locations of all such regions could help in identifying a common cause of low yield.

Linear regression analysis gives accurate results if all independent parameters are taken into account and the relationship between dependent and independent parameters is uniform across the field. The accuracy of linear regression analysis declines if we do not have sufficient information about all the independent parameters or they are uncontrollable. For example, in Figure 1, the points are well placed along the regression line and x and y follow a uniform relationship throughout. In such a case, a linear regression model gives an accurate relationship between x and y. Consider the real life scenario shown in Figure 2, which shows a relationship between yield (dependent variable) and NDVI (independent variable). In this plot, points are clustered and clearly not well placed along the regression line. There are various soil variables such as soil composition, water levels, etc. that affect this relationship. Some of these factors can be controlled while some of them cannot be. These external factors not only introduce errors but also make the relationship non-uniform across the field. Both, inaccuracies and variation in the relationship cannot be explained using one regression line. Figure 3 depicts such a case. Figure 3(a) depicts how the relationship could be varying at different parts of the field but one regression line models a completely different relationship as shown in Figure 3(b). We use the slope of regression line to denote the Relationship between x and y.

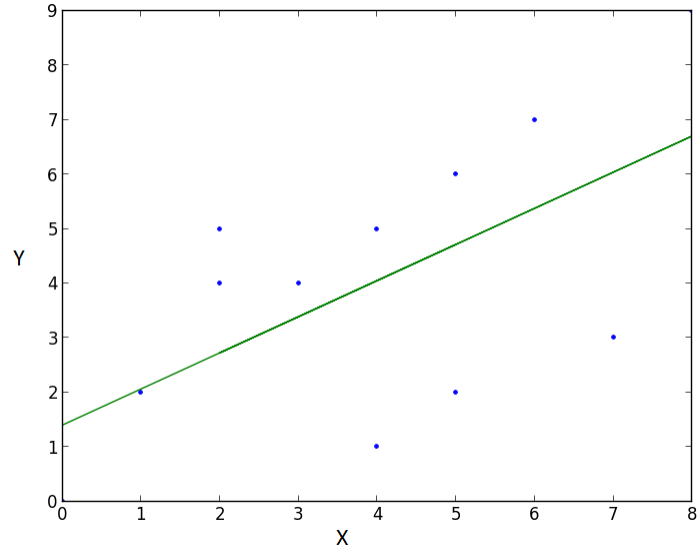


Figure 1. Ideal case for linear regression

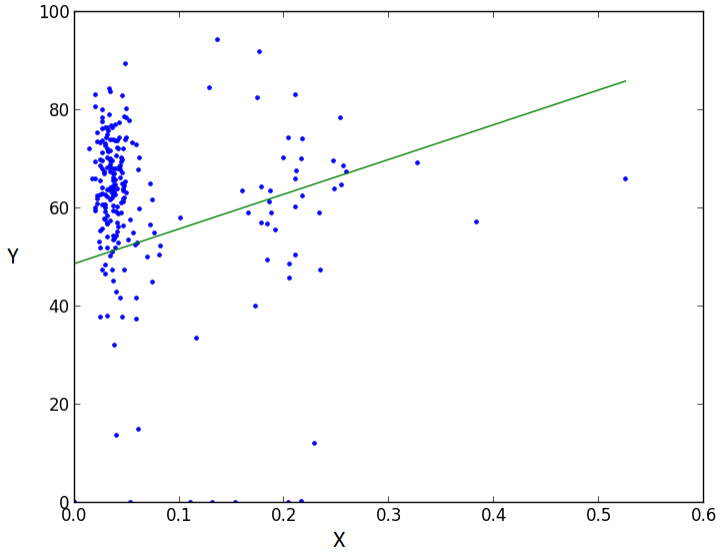
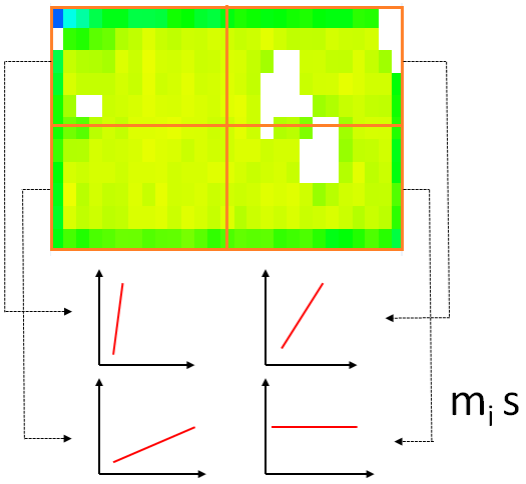


Figure 2. Real life scenario

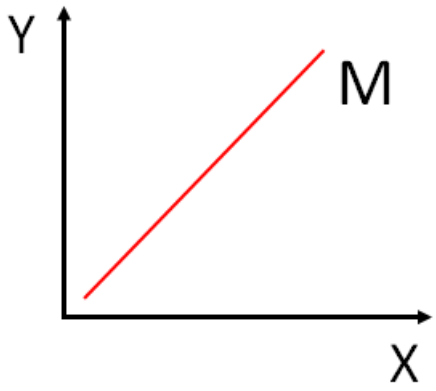
1.2. Problem Statement

The objectives of this work are as follows:

- To demonstrate that it is inaccurate and incomplete to represent correlation between different variables using a single regression line.



(a) Varrying relationship across the field



(b) Effective relationship depicted using one regression line

Figure 3. An example of inaccurate regression analysis using one regression line

- To propose the best way to represent a correlation between variables which do not follow a uniform relationship across the field.
- To select appropriate window sizes based on the evaluation of correlation over multiple window length scales.
- To demonstrate the importance of SWBT for farmers and to analyze various features over the field.

We propose a technique based on *sliding windows* to solve these problems. The key idea of our solution is to find relationship over small windows of field area instead of looking at the complete field at once. This significantly reduces the impact of the external factors on the correlation as those factors do no vary significantly over a small area. Chapter 3 explains the Sliding Window Based Technique (SWBT) in details along with the algorithm used and characteristic features.

Outline of thesis The next section describes the existing work related to our approach. We explain all the basic concepts needed to understand our technique in Chapter 2. Chapter 4 lists the experimental evaluation of SWBT. We conclude the thesis with conclusions and future work in Chapter 5.

1.3. Related Work

Numerous efforts have been made in the recent past to apply data mining techniques in agricultural-related fields. [9] summarizes some interesting applications of different data mining techniques such as k-means, *k*-NN classification, SVM, etc. in agriculture. Data mining techniques are often used to study soil characteristics. For example, [4] uses the k-means algorithm to find culsters of soils and plants. Predicting yield is a very attractive problem in agriculture [10], [11].

There are existing methods that make use of sliding window approaches for averaging purposes [1]. However, to the best of our knowledge, there are no reports of using window based techniques the way we used it, in agriculture. [14] is an example from the field of image processing that uses windows for adaptive images contrast enhancement. They compare the grey level of a pixel with the neighboring pixels (pixels in its window) and decide the output grey level. Window based techniques are heavily used in the field of data stream processing. The problem of creating fast histograms for sliding window streams and time series is a research topic by itself [12], [13]. They maintain the data statistics over small windows and make use of them while building the sub-optimal histograms for the complete data.

CHAPTER 2. CONCEPTS

In this chapter, we explain the basic concepts required to understand this work.

2.1. Satellite Imagery

We use satellite images of agricultural fields as data source. These images are taken by Landsat 8 satellite [15] and they are freely available at [16]. The Landsat 8 satellite images the entire Earth every 16 days and collects data from nine spectral bands. We make use of Band 4 (Red) and Band 5 (Near Infrared). These images are in the raster format. The resolution of these images is 30 m. Therefore, each pixel represents data for $900m^2$ area on the field. We also make use of high resolution images taken from commercial satellite [2]. The resolution of these images is 5m.

2.2. Normalized Difference Vegetation Index (NDVI)

NDVI is an index that assesses whether the target being observed contains live green vegetation or not. It is an indicator of the green-ness of the field. NDVI is calculated from the spectral reflectances in the Red and Infra-Red bands. The mathematical formula for NDVI is:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

where, RED and NIR stand for the spectral reflectance measurements acquired in the visible (red) and near-infrared regions, respectively. The value for NDVI ranges from 0 to 1. Higher the NDVI higher is the vegetation. For this work, we make use of raster images from [2] consisting of NDVI information. An example of rasterized NDVI image is shown in Figure 4.

2.3. Yield

Yield is the amount of crop harvested per unit of land area. Sensors are attached to the harvesters to collect yield information. On getting the yield information from farmers, we convert

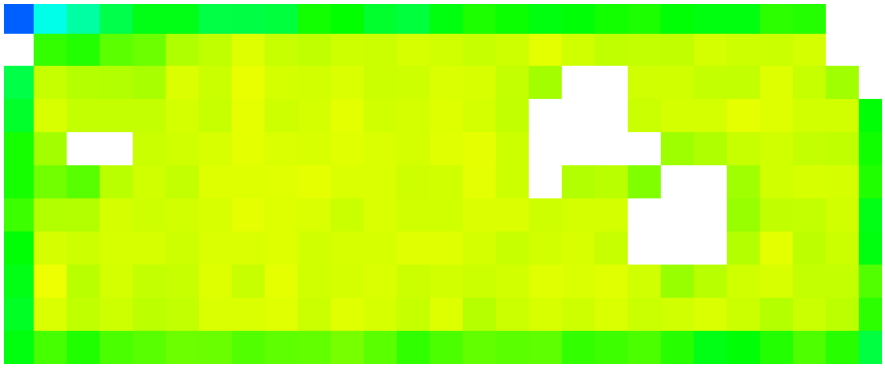


Figure 4. Sample NDVI image

it to raster format and use the raster images for data processing. An example of rasterized yield image is shown in Figure 5.

Figure 5. Sample yield image

2.4. Scatter Plot and Linear Regression

A Scatter plot is a plot of the values of the dependent variable (on Y axis) versus the independent variable (on X axis). Linear regression is a statistical technique to model the relationship between dependent and one or more independent variables. Where there is only one independent variable, it is called as simple linear regression. In this work, we only consider simple linear regression. This relationship between variables is indicated by a regression line, which is used to predict the values of the dependent variable or to quantify the strength of relationship between the dependent variable and one or more independent variables. Figure 6 shows an example of a scatter plot and regression line.

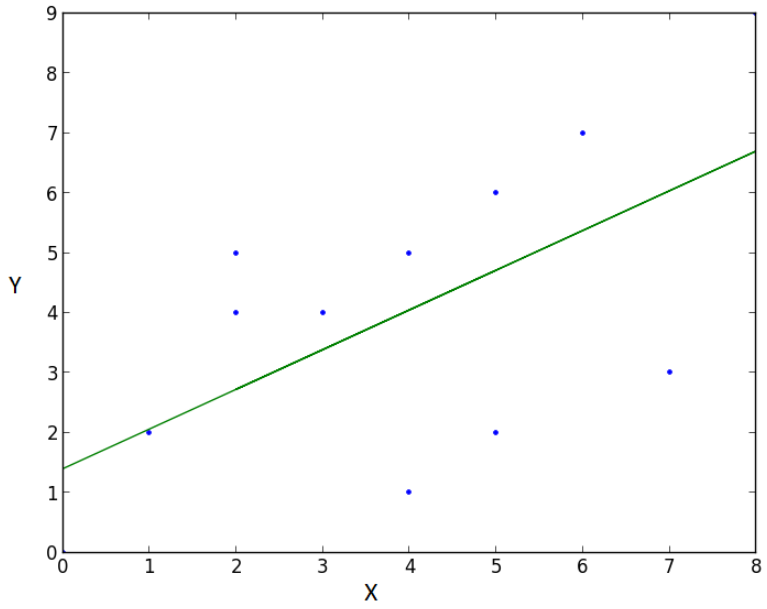


Figure 6. Sample scatter plot and regression line

In the above example, X={0,1,2,3,4,5,6,7,8,2,4,5} and Y={0,2,5,4,1,2,7,3,9,4,5,6}. The equation of the regression line giving relationship between X and Y is: y=0.66x+1.4 where 0.66 is the slope of the line and 1.4 is the intercept.

CHAPTER 3. SLIDING WINDOW BASED TECHNIQUE

In this chapter, we explain SWBT along with its characteristic features and the algorithm. The key idea of SWBT is localization. Instead of looking at the complete field for determining the relationship between variables, we propose to determine the relationship over several small areas of the field. Such areas of the field are called windows. We consider square shaped windows, however, they can be of any shape. Because the input data files are rasterized, we read the field information pixel by pixel. Every window is composed of several pixels and is identified by the pixel on the left top corner. A window of size 3 contains $3 * 3 = 9$ pixels. These windows are overlapping, meaning every window (except the ones on the field boundary) differs by a row of pixels from above and below windows and a column of pixels from left and right windows. The pictorial representation of these windows is given in Figure 7 where different overlapping windows are shown with different colors.

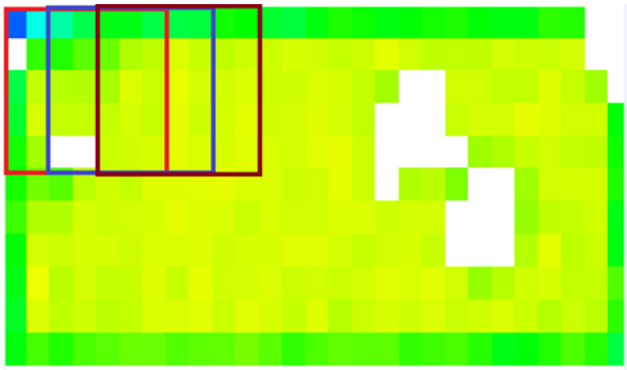


Figure 7. Respresentation of sliding windows

As we progress from left to right, windows move by a column of pixels. And as we move from top to bottom they move by a row of pixels. Therefore, these windows are called as sliding windows.

3.1. Characteristic features of SWBT

Here are the characteristic features of SWBT.

- Within a window, due to localization, external factors do not change significantly resulting in less impact of external factors on the correlation. Therefore, the results obtained in each window are significantly more accurate.
- Within each window we calculate the correlation using linear regression where the correlation is represented by the slope of the regression line. Therefore, in the end we get a slope distribution across the field.
- We also get an idea of how the slope varies over the complete field, which is essentially how the correlation changes over the complete field.
- Clustering of slope can be used to find regions of similar slopes.

3.2. Algorithm

Algorithm 1 lists the steps involved in SWBT. It shows an example of how the algorithm works while calculating the relationship between yield and NDVI. Using the notations given in the algorithm, there would be a total of (endX-startX-W+2) * (endY-startY-W+2) windows.

3.2.1. Window-size selection

We evaluate Algorithm 1 for multiple values of W. For selecting appropriate values of W, we plot a histogram of the window-ised slopes (m_is) and identify multiple maxima in the histogram. We select only those window sizes that produced multiple maxima as appropriate values of W. Multiple maxima represent regions on the field with different yield levels and the minimum between the peaks indicates the boundary between the region in the field in which the yield does not vary even when the NDVI varies, and the region where there is a substantial dependence of yield

Algorithm 1 Sliding Window Based Technique

Input

startX, *startY*: X and Y co-ordinates of the top left corner of the field image
endX, *endY*: X and Y co-ordinates of the bottom right corner of the field image
W: Sliding window size
yield: Raster file containing yield data
ndvi: Raster file containing NDVI data

Output

m_i: An array of slopes (of regression lines) for each sliding window

```
1:  procedure SWBT
2:
3:  i=startY, j=startX
4:      for each i <endY do
5:          for each j <endX do
6:              row=0, column=0
7:              Initialize arrays X and Y to null
8:              for each row <W do
9:                  for each column <W do
10:                     Append ndvi.getpixel((j+column, i+row)) to array X
11:                     Append yield.getpixel((j+column, i+row)) to array Y
12:                 end for
13:             end for
14:             Calculate slope using X and Y arrays:
15:             slope = numpy.polyfit(X, Y, 1)
16:             Append slope to output array m_i
17:         end for
18:     end for
19:
20: end procedure
```

on NDVI. To clearly identify maxima in the histogram, we smooth out the histogram. Smoothing is done by calculating the sum of frequencies over neighboring 15 bins of histogram and replacing it for the frequency value of the middle bin. We further use selected window sizes to identify regions of high and low yield on the field.

CHAPTER 4. EXPERIMENTAL RESULTS AND OBSERVATIONS

In this chapter, we describe the implementation details of SWBT. We also describe the experiments done to evaluate SWBT and observations pertaining to every experiment.

4.1. Implementation Details

In this section, we explain the implementation details of SWBT.

- **Data** We used satellite images [16] from Landsat 8 satellite for NIR and NER data, high resolution satellite images from Satshot [2] for NDVI and data from farmers for yield (bushels/acre) information. The data belongs to the year 2012 for a wheat field in Stutsman County, ND. The NDVI images were taken on June 22, 2012.

- **Data cleaning and conversion** We used GRASS GIS [3] for data cleaning and conversion purposes. GRASS GIS, commonly referred to as GRASS (Geographic Resources Analysis Support System), is a free and open source Geographic Information System (GIS). As said earlier, we received yield data in vector file format and NDVI, NIR, NER data in raster format. It was necessary to use a common data format. The projection format of the NDVI raster image was "WGS84" and that of yield data was Lat-Long. We used GRASS command-r.proj to re-project NDVI raster file into Lat-Long projection format. To convert vector data into raster format, we used GRASS command v.to.rast for the attribute DryYield. Once both NDVI and yield data files were in same projection and format, we exported them as TIF files for better readability through Python programs.

- **Data processing** The programs for data processing were implemented in Python. Python Image Processing Library is used to read the TIF files pixel by pixel and process the data. Matplotlib, a Python 2D plotting library is used to display scatter plots, regression lines and histograms.

For high resolution images, the size of the field under consideration was 56 pixels (rows) X 144 pixels (columns) i.e. 16 hectares. In all of the linear regression plots, the independent variable was NDVI and the dependent variable was yield, unless mentioned otherwise.

4.2. Experiments

4.2.1. Slope distribution across the field

In order to demonstrate that a single regression line is not enough to represent the correlation between parameters, we found out the window-ised slope distribution across the field and compared it with the slope of the line when a single regression line is used for the complete field. m_i represents the window-ised slopes and M represents the slope of the regression line when only one regression line is used to represent the correlation across the complete field. Figure 8 shows the histogram of M versus m_is for window size=15, where on the X axis are the various intervals of slopes- *bins* and Y axis indicates the count of slope values falling in those bins- textitfrequency. The bins for window-ised slopes are indicated in blue color and the overall slope is indicated in green color. In reality, there is only one value of M, however for the ease of comparison it is shown higher to make the bin taller in the histogram. Table 1 compares the data statistics related to the values of m_is and M.

Table 1. Results for slope distribution across the field for W=15

M	W=15		
	Mean of m_is	Median of m_is	Range of m_is
154.53	65.33	109.38	[-727.712, 806.29]

From Table 1, it can be observed that the mean and median of m_is is significantly different than M. For a smaller area such as W=15, we can see the benefits of localization which make the relationship between Yield and NDVI accurate within every window. Because of such small area, the values being aggregated are less, therefore, each window tends to be different than others which

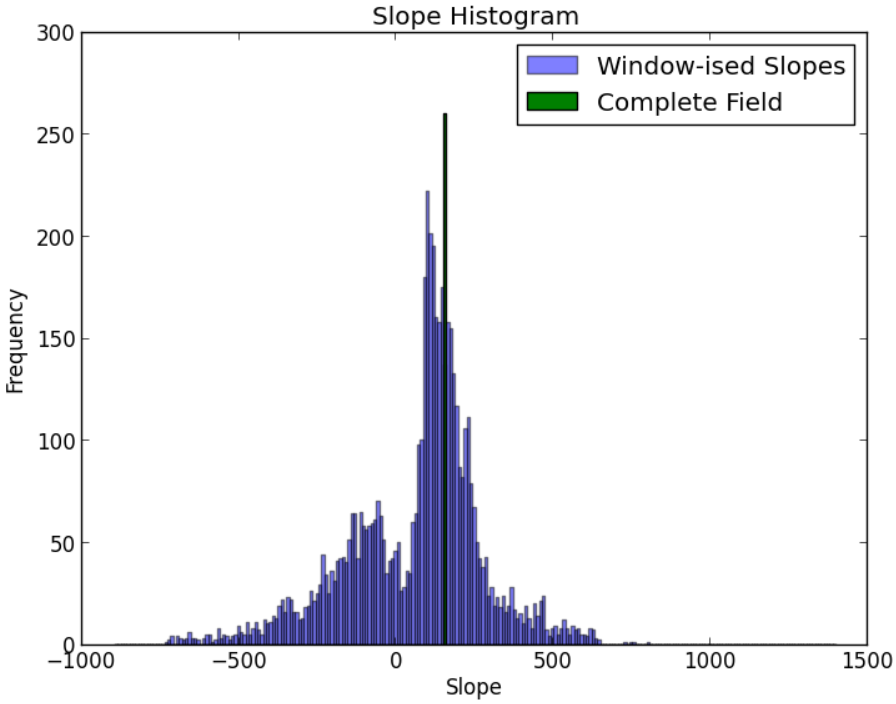


Figure 8. Slope distribution across the field (W=15)

leads to a wide range of m_is indicating that yield and NDVI do not follow uniform relationship over the complete field.

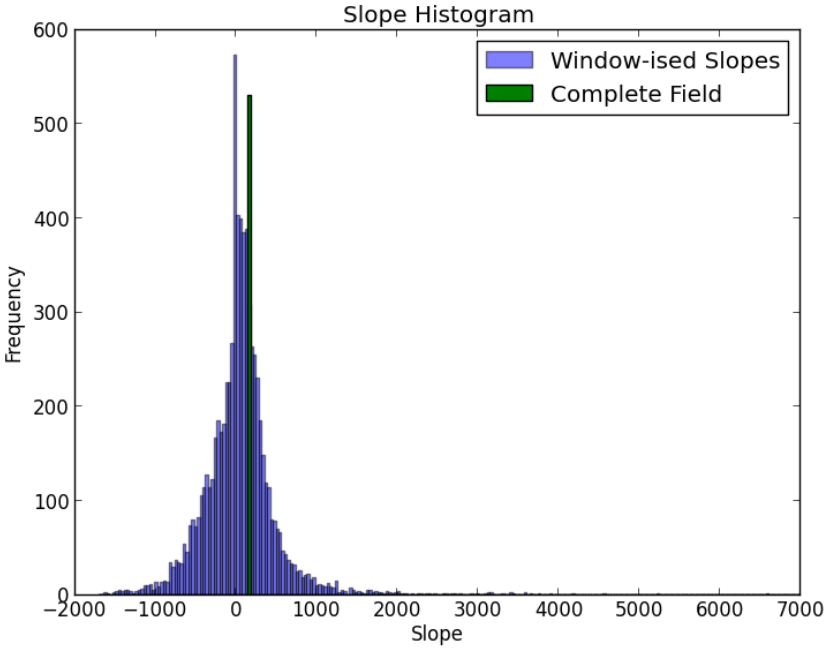
In many ways, having a histogram representation of m_is is more useful than having a single value of slope for the entire field. Histogram representation depicts the trends in the relationship across the whole field. It gives us finer details about the relationship. With the help of maxima and minima present in the histogram we can find the regions of low yield-predictability and high yield-predictability on the field. The range of m_is indicates the variability in the relationship.

4.2.2. Slope distribution with different window sizes

In order to see if the slope distribution is a function of sliding window size (W), we performed a similar experiment for different window sizes (W). Figure 9 shows all the histograms for different values of W and Table 2 lists data statistics.

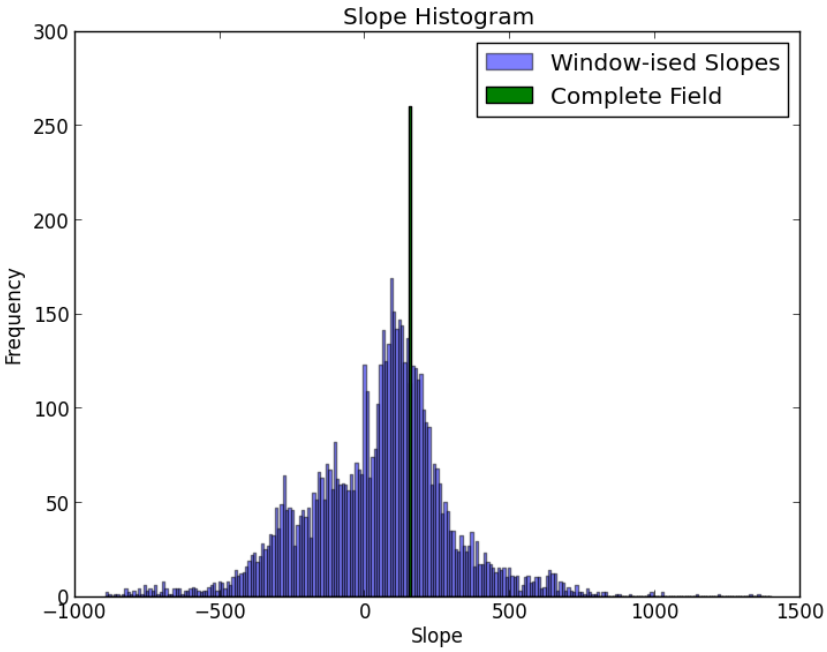
Table 2. Results for slope distribution across the field for different values of W

W	M	Mean of m_is	Median of m_is	Range of m_is
5	154.53	48.79	40.71	[-1704.03, 6587.25]
10	154.53	46.19	76.04	[-888.17, 1359.30]
15	154.53	65.33	109.38	[-727.71, 806.29]
20	154.53	91.14	125.38	[-475.05, 513.65]
25	154.53	112.05	136.47	[-368.30, 327.62]
50	154.53	132.66	146.23	[-23.43, 225.98]
55	154.53	137.98	151.21	[-5.32, 213.06]
56	154.53	139.08	150.86	[-2.78, 209.57]

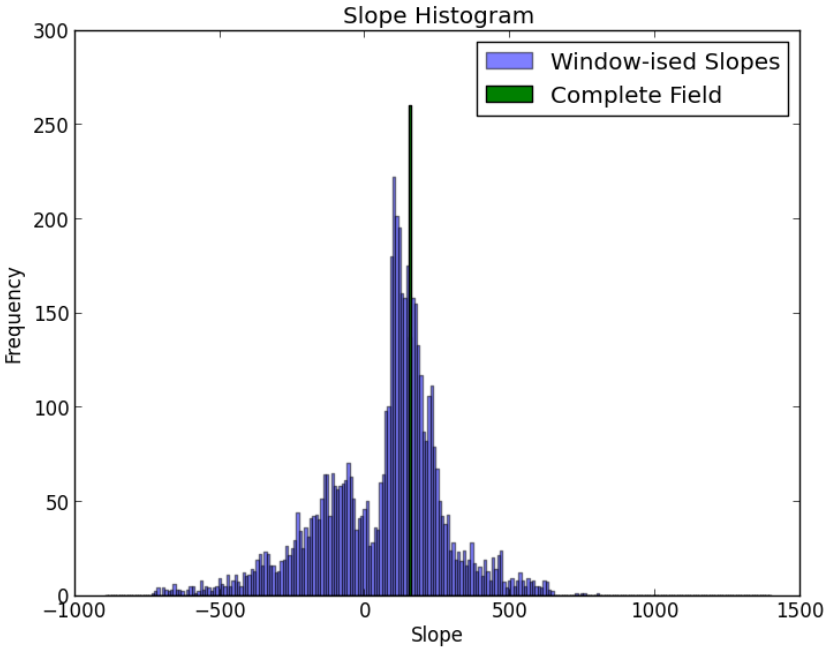


(a) W=5

Figure 9. Slope distribution for different values of W

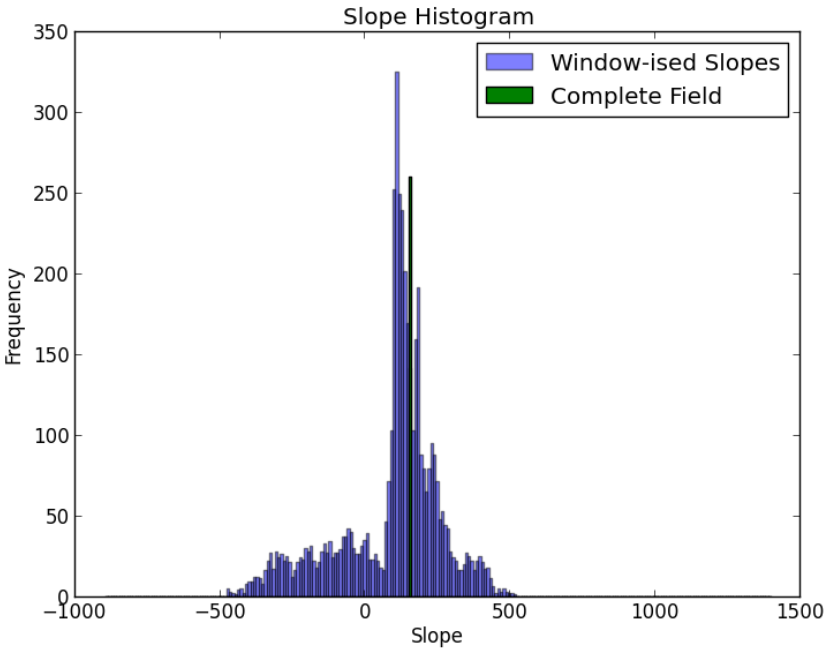


(b) W=10

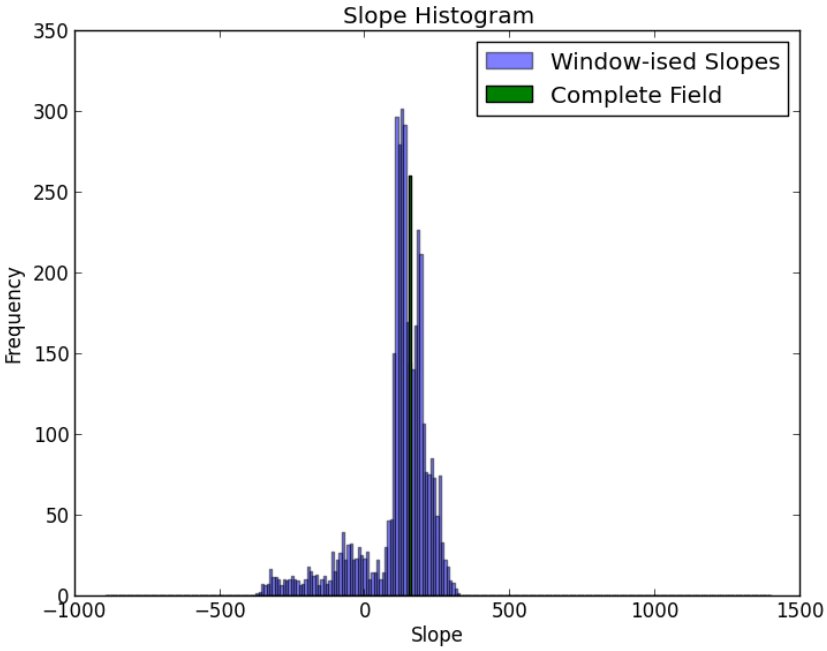


(c) W=15

Figure 9. Slope distribution for different values of W (continued)

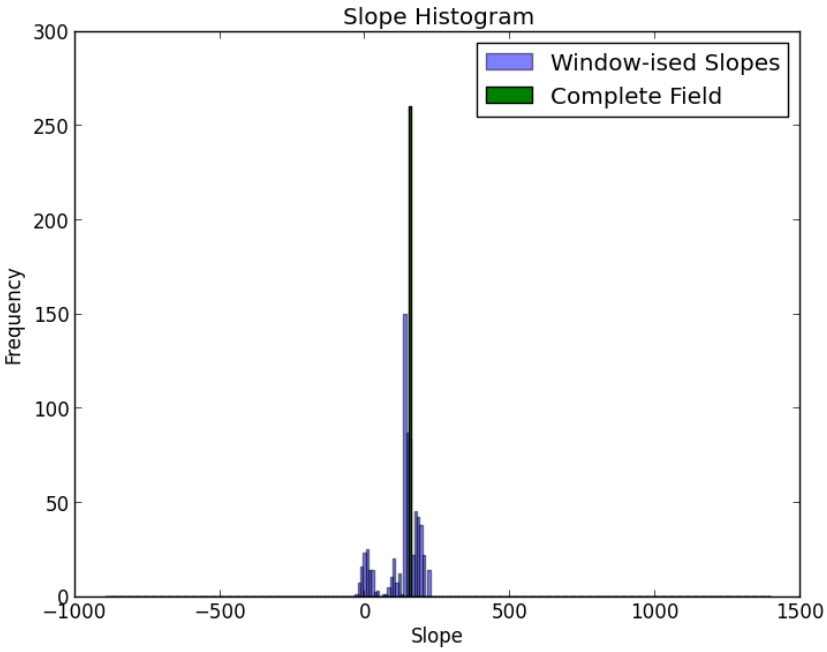


(d) W=20

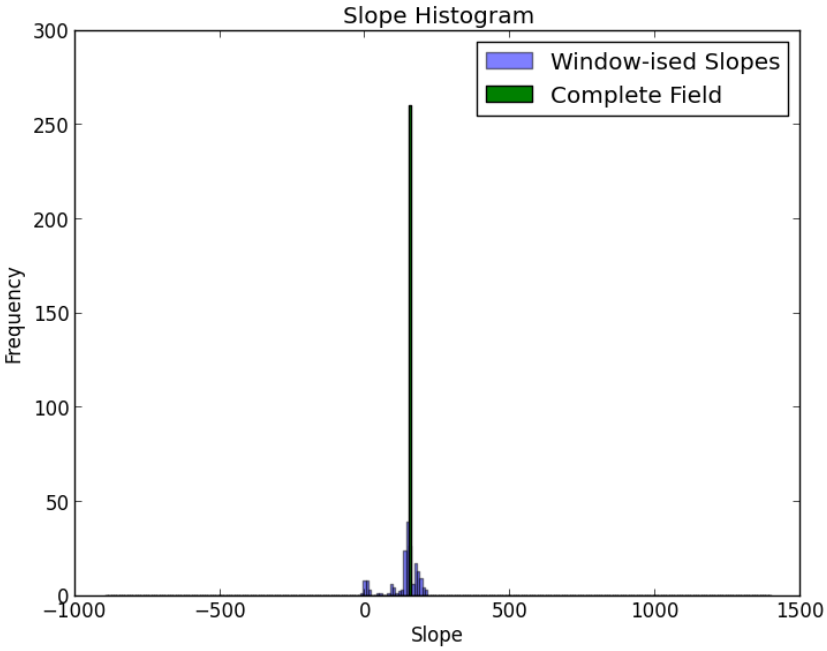


(e) W=25

Figure 9. Slope distribution for different values of W (continued)

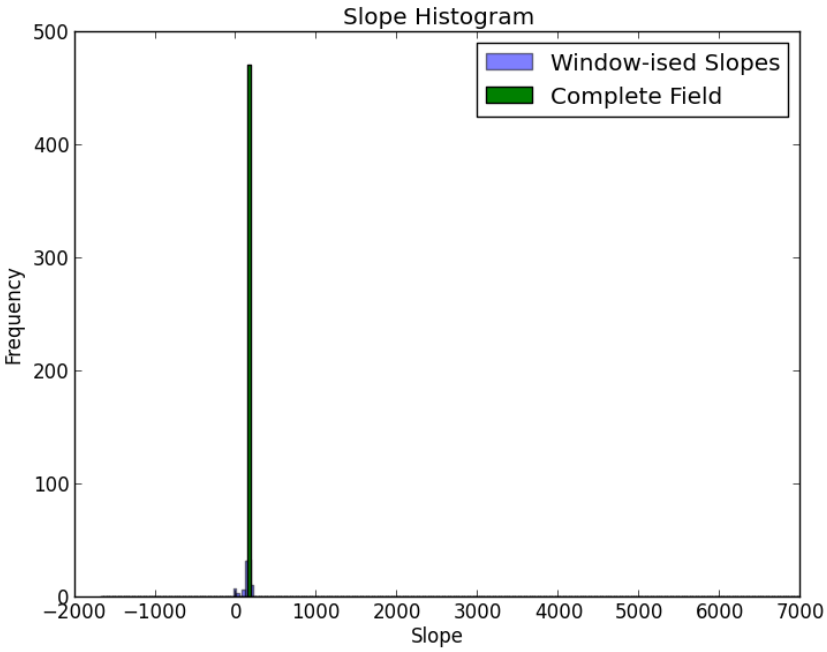


(f) W=50



(g) W=55

Figure 9. Slope distribution for different values of W (continued)



(h) W=56

Figure 9. Slope distribution for different values of W (continued)

For small window sizes such as W=5, 10, 15, 20, each window consists of a small number of pixels- 25, 100, 225, 400, respectively. Each slope value represents the aggregation over a small area, therefore, less averaging occurs for X and Y values, resulting in a wide range of m_is. Also, because modelling the relationship in smaller windows is more accurate the median/average of window-ized slopes is a lot different than the slope of the regression line considering the complete field as one window. As W increases, each window spans over larger a field area, thus, aggregating more values. Therefore, due to averaging of X and Y values, windows start to have similar slopes resulting in small range of slope values. Due to variability in soil contents, the accuracy of correlation declines which brings the mean/median of m_is closer to M. At W = 56, the median slope is almost equivalent to M.

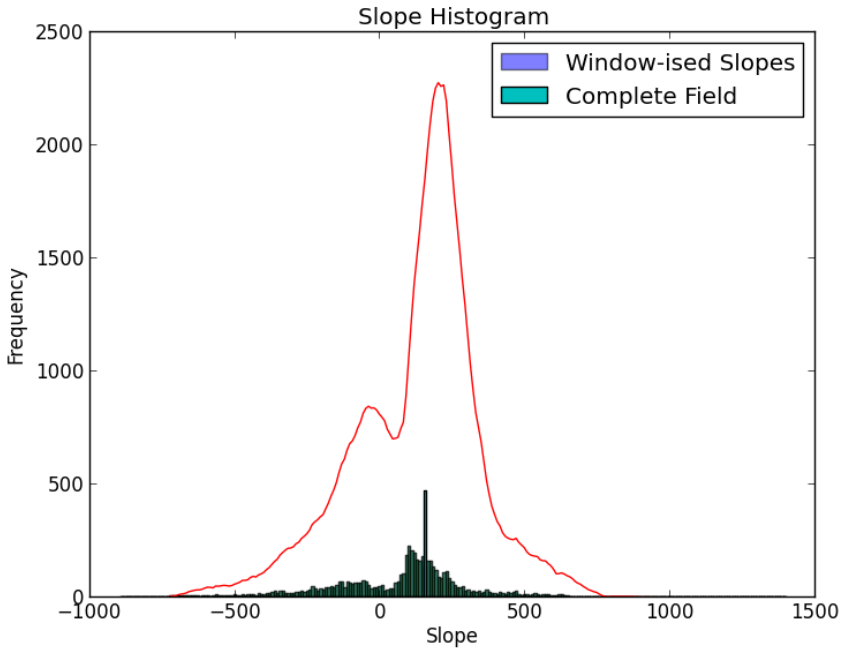
As shown in Figure 9(a), the range of slopes for W=5 is [-1704.03, 6587.25] and the median of m_is is 40.71 which is significantly lower than M (154.53). In Figure 9(b), the range of slopes

decreases drastically to [-888.17, 1359.30] as W increases from 5 to 10. The median of m_is also jumps up to 76.04. The range of slopes goes on decreasing with further increase in the value of W (Figure 9(c), 9(d), 9(e), 9(f) and 9(g)). As we can see in Figure 9(h), the range of m_is drops to [-2.78, 209.57] for W=56 when the median of m_is is 150.86, which is almost similar to the value of M.

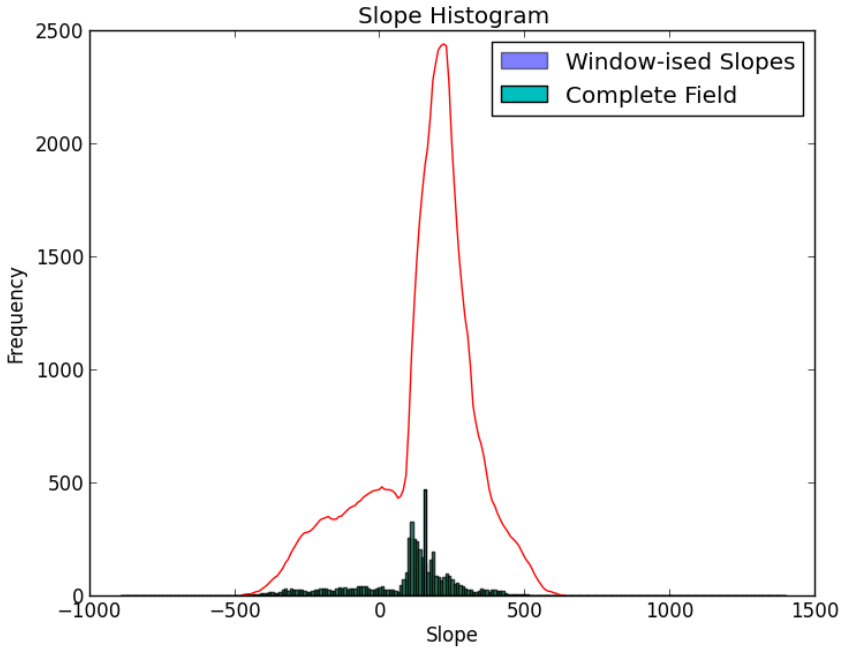
Both these experiments validate our claim that having a single regression line does not give accurate and complete correlation between variables. And such a non-uniform relationship can be best represented by a slope distribution histogram.

4.2.3. Window-size selection

In Figure 9(c), 9(d) and 9(e), for W= 15, 20 and 25, respectively, we can see multiple maxima in the histograms. In order to find the slope values separating these two peaks, we performed smoothing of the histograms. Figure 10 shows the histograms for W=15, 20 and 25 after smoothing. For smoothing, we calculated the sum of frequencies over neighboring 15 bins of the histogram and replaced it for the frequency value of the middle bin.

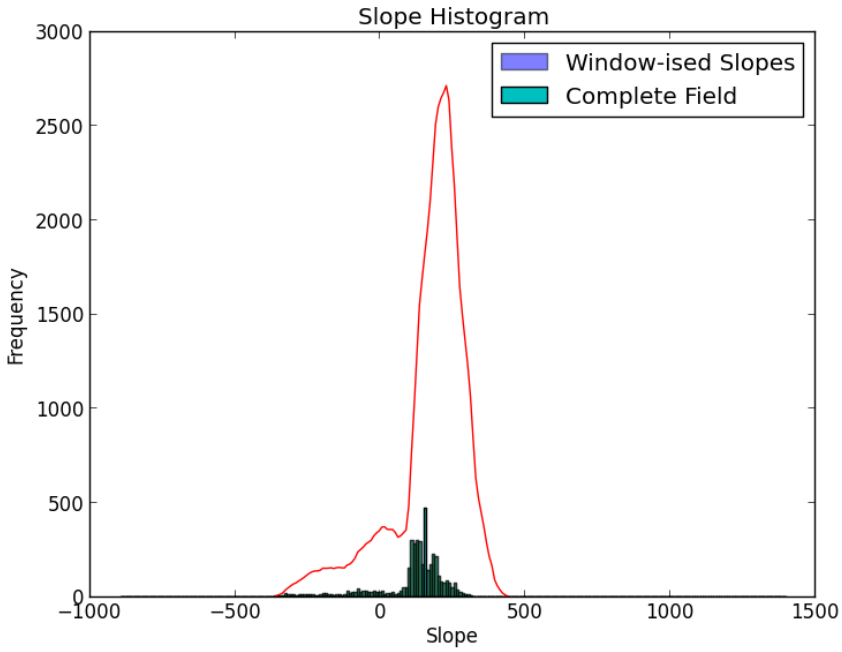


(a) For W=15



(b) For W=20

Figure 10. Slope distribution histograms after smoothing



(c) For W=25

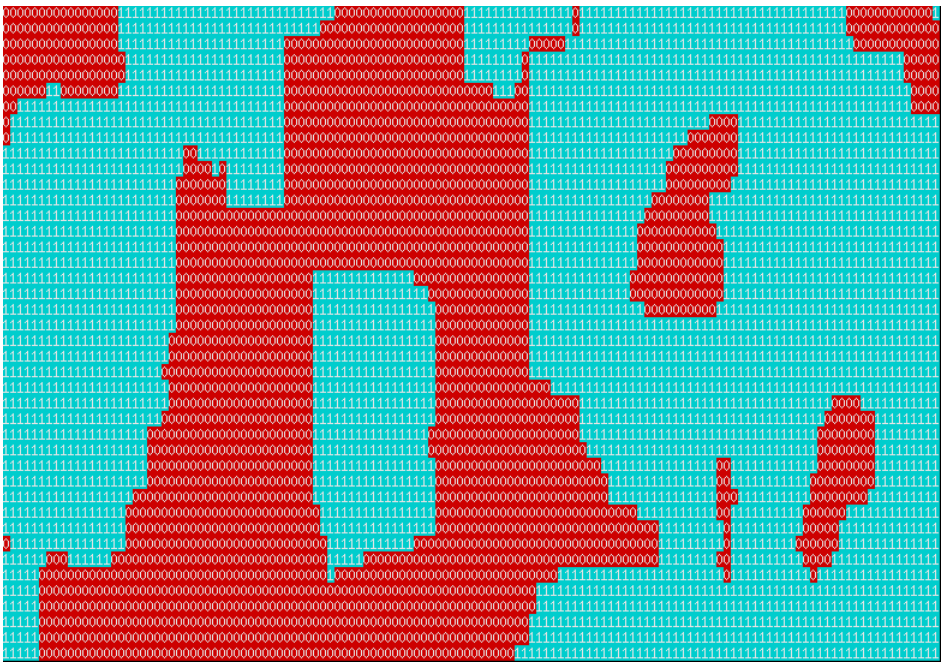
Figure 10. Slope distribution histograms after smoothing (continued)

Figure 10 shows 2 maxima that are clearly separated. The slope values corresponding to the minimum between the two maxima is almost identical in all three cases. Since the slope is an indication of the relationship between yield and NDVI, the minimum indicates the boundary between the region in the field in which the yield does not vary even when the NDVI varies, and the region where there is a substantial dependence of yield on NDVI. We select these three window sizes for further analysis.

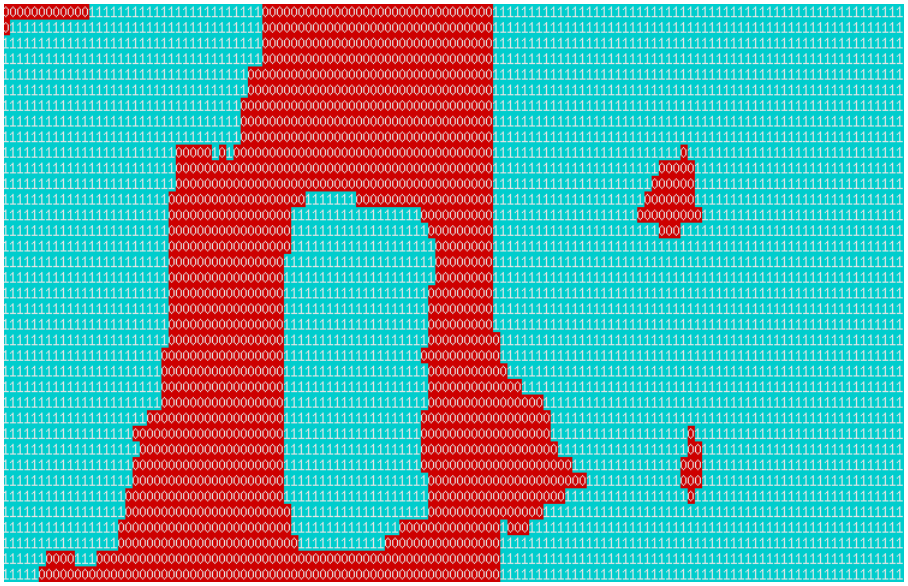
4.2.4. Clustering of slopes to find the regions of low yield-predictability and high yield-predictability

As indicated in Figure 10, using the boundary value belonging to the ditch, we can cluster the field into two regions- one with a high yield-predictability and one with a low yield-predictability. Figure 11 indicates these two clusters for all three values of W. We used hard partitioned clusters

with 85 bushels/acre as the boundary slope. The region in red color (indicated by 0s) is the region with low yield-predictability, and the region in cyan color (indicated by 1s) is the region with high yield-predictability.

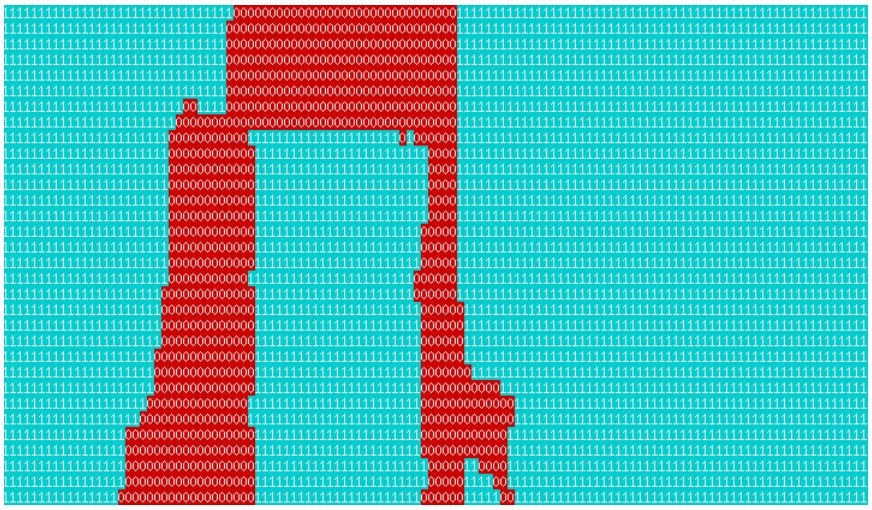


(a) For W=15



(b) For W=20

Figure 11. Clusters of low and high yield over the field



(c) For W=25

Figure 11. Clusters of low and high yield over the field (continued)

At small values of W (W=15), the clusters indicate minute details about the low yield-predictability regions and high yield-predictability regions. As the value of W increases, windows start to span over a large area on the field, aggregating more values. Therefore, the slope values start to smooth out. As a result, clusters with high values of W do not show fine details. For example, in Figure 11(a) there are many small low yield-predictability regions indicated. We can clearly see a low yield-predictability region on the left top and one on right top corner, which start diminishing as we increase the value of W. In Figure 11(b), we can see a small region on the left top corner but the region on the right top corner is not characterized as low yield-predictability. Both these regions completely disappear in the clusters shown for W=25 in Figure 11(c).

These clusters along with the soil map shown in Figure 13 can be used to find out the relationship between soil map and yield-predictability of the soil type. It can be seen from Figure 11 that soil types G143B, G101A have high yield-predictability whereas, soil types G143C, G144B have low yield-predictability.

Such analysis can help farmers identify the regions of low yield and they can take corrective measures (applying fertilizers) to improve yield-predictability in such regions. Studying the loca-

tion of the low yield-predictability regions, we can determine a common cause of low yield. For example, if all the low yield regions are indicated near water bodies that means the water content in the soil near such areas is higher than what is ideal for high yield. Farmers can make use of such specific information to take corrective actions.

4.2.5. Analysing clusters of slopes

As explained earlier, the variability in soil composition, water levels, etc are some of the factors behind the non-uniform nature of the relationship. Unpredictable changes in these external factors cause unpredictable change in the relationship trend. We performed an experiment to observe if the relationship between yield and NDVI is a function of soil type.

The soil maps for fields are freely available at [7]. Figure 13 indicated the soil map for the field under consideration. As we can see, there are a total of five different types of soils present in the field. This categorization of soils is based on the vegetative index. We used the values of m_is calculated using Algorithm 1 explained in Chapter 3. We formed clusters of these values using an existing implementation of the *K*-means algorithm in Python with k = 5 and compared them with the soil map of the field soil map. Figure 12 indicates different clusters of slopes with different colors.

From the overlay of Figure 12 and Figure 13, we tried to find out if there is a similarity in the shape of soil map and the clusters. However, as you can see, the clusters do not follow the soil map. It can be observed that there is a many-to-many relationship between soil type and slope clusters. Meaning, one soil type consists of multiple clusters and slope values belonging to one clusters can exist in multiple soil types. For example, soil type G143B contains all five clusters and parts of the cyan cluster can be observed on soil types G143B, G143C, G144B and G101A. Therefore, it is clear that the relationship between yield and NDVI is independent of soil type. Also, it is non-uniform even within one soil type.

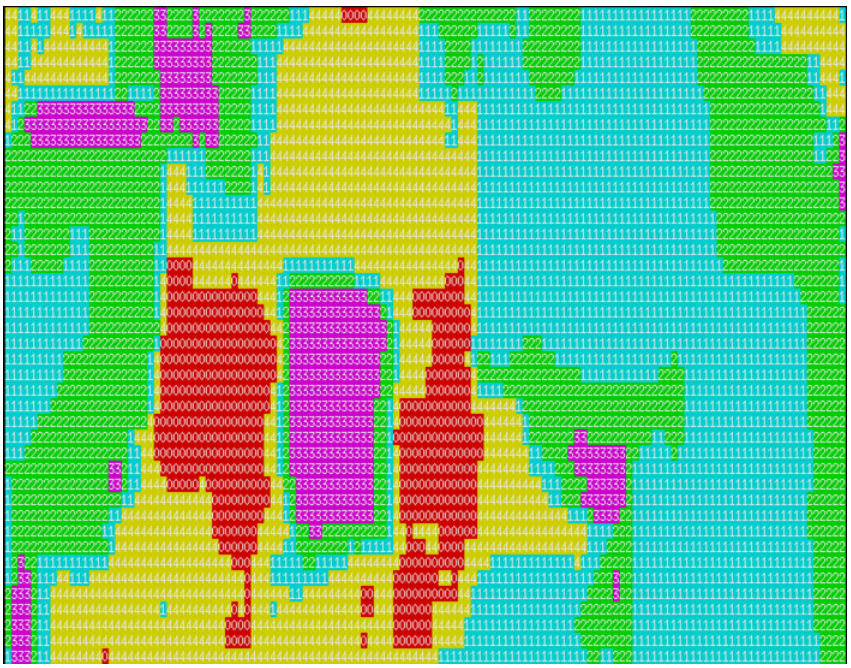


Figure 12. Clusters of slopes (W=15)

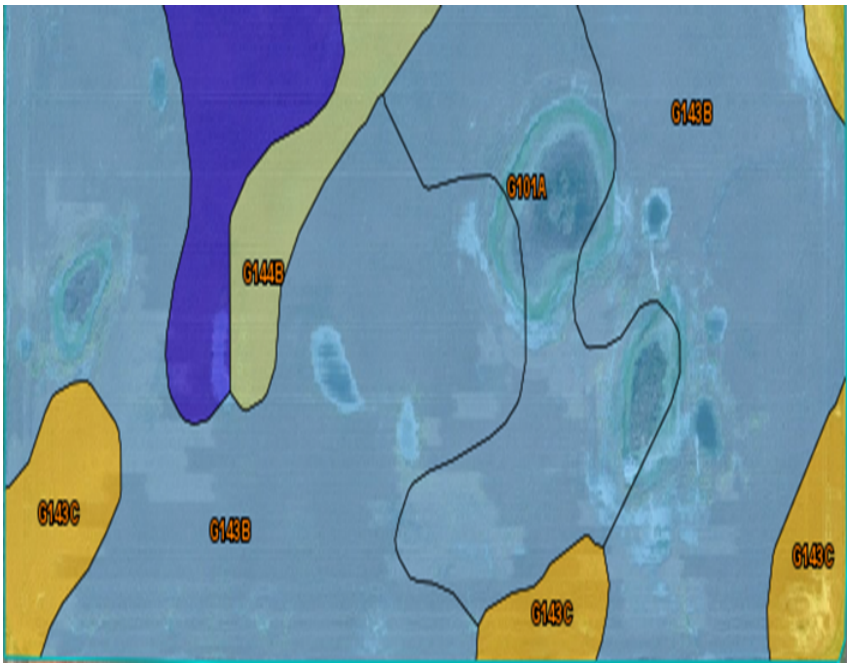


Figure 13. Soil map for the field

4.2.6. Classification between field and non-field regions

In order to find the accuracy of SWBT, we performed an experiment to figure out how well SWBT can classify between field and non-field regions. For this experiment, we took into consideration an area which had field and non-field regions. Instead of yield Vs NDVI, we used the relationship between NIR (dependent variable) and NER (independent variable) to distinguish between field and non-field regions because yield data is only applicable to fields. NIR and NER data is available from the Landsat satellite. The resolution of this data was 30m. Therefore, every pixel in the image represented an area of $900m^2$ on the field. The field size in terms of pixels was 24 (rows) X 25 (columns). The size of non-field region was also 24 (rows) X 25 (columns). We executed Algorithm 1 on these regions and plotted the histogram. The window size was 4. Figure 14 shows the histogram of slopes belonging to field and non-field regions. Slope bins belonging to the field region are indicated in blue color and slope bins belonging to the non-field region are indicated in green color. Table 3 lists the statistics related to histograms.

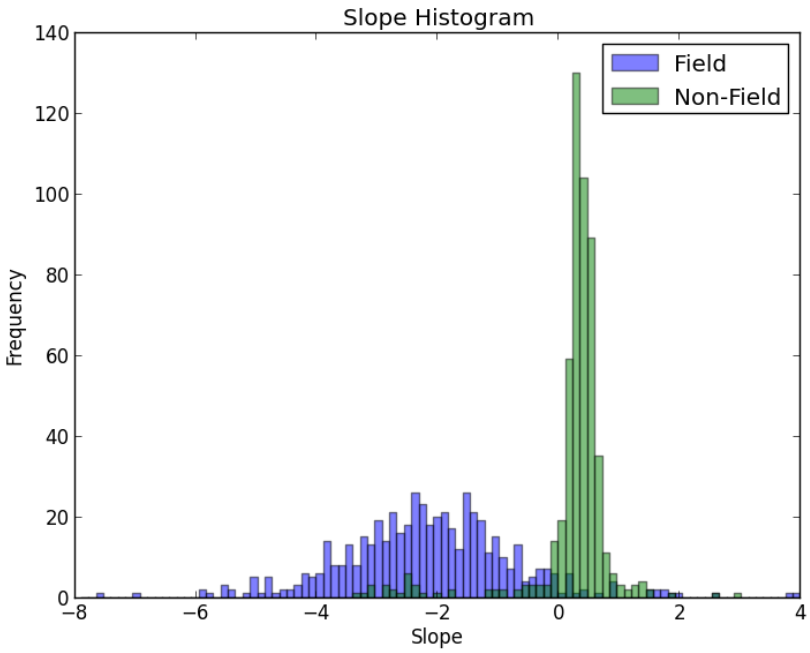


Figure 14. NIR Vs NER to distinguish between field and non-field regions

Table 3. Statistics for slopes (NIR Vs NER) belonging to field and non-field regions

	Field Region	Non-field Region
Mean of m_is	-2.15	0.25
Median of m_is	-2.18	0.36
Range of m_is	[-7.60, 3.92]	[-3.32, 3.02]

It can be observed from Table 3 that the mean and median slopes of field region are significantly different than the mean and median slopes of non-field region. The mean slope of field was -2.15 whereas the mean slope of non-field region was 0.25. The median slopes in case of field and non-field regions were -2.18 and 0.36 respectively, which were close to respective mean slopes but were, again, significantly different than each other. The range of field slopes- [-7.60, 3.92] was wider than that of non-field slopes- [-3.32, 3.02]. The ranges were not exclusive, however, as seen in Figure 14, the dense parts of both histograms were distinctly separable from each other. This indicates that the SWBT using NIR and NER attributes can be used to distinguish between field and non-field regions on the ground.

It is possible to get the exact accuracy using any one of the classification algorithms with the help of number of true positives, true negatives, false positives and false negatives.

CHAPTER 5. CONCLUSIONS AND FUTURE WORK

With the increasing availability of huge volumns of data in agriculture, we need advanced data mining techniques to process this data efficiently. We proposed a novel technique- Sliding Window Based Technique, to calculate the correlation between various parameters of the field. In order to model a non-uniform relationship it is always beneficial to make use of localized information. We worked with small areas of field-windows to find the correlation between parameters. We demonstrated that this way of finding correlation is more accurate and complete than representing the correlation with only one model for the complete field. We also demonstrated the non-uniform nature of relationship across the entire field due to external variables. Using multiple window sizes we demonstrated how the relationship varies across the field. We also presented a method for selecting the appropriate window. With the help of peaks in the slope distribution histograms and clustering we presented how SWBT can be used to find low yield-predictability and high yield-predictability regions of the field. Farmers can make use of this information to improve the yield. Using clustering and soil map, we proved that the relationship between yield and NDVI is independent of soil type. We demonstrated that SWBT can be used to distinguish between field and non-field regions.

A possible extension to this work would be to further optimize the algorithm and integrate the evaluation over multiple window sizes into the sliding-window generation process.

REFERENCES

[1] *From land cover to landscape diversity in the european union*, Available at http://ec.europa.eu/agriculture/publi/landscape/, Accessed: 05-08-2015.

[2] *Satshot*, Available at https://www.satshot.com/, Accessed: 4-30-2015.

[3] *GRASS GIS*, Available at http://grass.osgeo.org/, Accessed: 4-30-2015.

[4] G. Meyer, J. Neto, D. Jones and T. Hindman, *Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images*, Computers and Electronics in Agriculture **42** (2004), 161 – 180.

[5] D. Lobell, *The use of satellite data for crop yield gap analysis*, Field Crops Research **143** (2013), no. 0, 56 – 64, Crop Yield Gap Analysis Rationale, Methods and Applications.

[6] W. Marshall, *Tiny satellites show us the earth as it changes in near-real-time*, TED2014.

[7] United States Department of Agriculture, *U. S. general soil map*, Available at http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/, Accessed: 4-30-2015.

[8] A. Mucherino, P. Papajorgji, and P. Pardalos, *Data mining in agriculture*, Springer-Verlag New York, 2009.

[9] A. Mucherino, P. Papajorgji and P. Pardalos, *A survey of data mining techniques applied to agriculture*, **9** (2009), 121 – 140.

[10] R. Medar and V. Rajpurohit, *A survey on data mining techniques for crop yield prediction*, **2** (2014), 59 – 64.

[11] G. Ruß, *Data mining of agricultural yield data: A comparison of regression models*, (2009), 24 – 37.

[12] S. Guha and N. Koudas, *Approximating a data stream for querying and estimation: algorithms and performance evaluation*, 18th International Conference on Data Engineering **18** (2002), 567–576.

[13] S. Guha, P. Indyk, S. Muthukrishnan and M. Strauss, *Histogramming data streams with fast per-item processing*, Proceedings of ICALP (2002), 681–692.

[14] A. Stark, *Adaptive image contrast enhancement using generalizations of histogram equalization*, IEEE Transactions on Image Processing **9** (2000), 889–896.

[15] United State Geological Survey, *Landsat 8*, Available at http://landsat.usgs.gov/landsat8.php, Accessed: 4-30-2015.

[16] United State Geological Survey, *Landsat project*, Available at http://landsat.usgs.gov/, Accessed: 4-30-2015.

[17] A. Chinchuluun, P. Xanthopoulos, V. Tomaino and P. Pardalos, *International journal of agricultural and environmental information systems (ijaeis)*, International Journal of Agricultural and Environmental Information Systems (IJAEIS) **1** (2010), 26 – 40.