INCORPORATING SLIDING WINDOW-BASED AGGREGATION FOR EVALUATING TOPOGRAPHIC VARIABLES IN GEOGRAPHIC INFORMATION SYSTEMS

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

By

Rahul Gomes

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

August 2019

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

INCORPORATING SLIDING WINDOW-BASED AGGREGATION FOR EVALUATING TOPOGRAPHIC VARIABLES IN GEOGRAPHIC INFORMATION SYSTEMS

By

Rahul Gomes

The supervisory committee certifies that this dissertation complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Anne. M. Denton
Chair

Dr. Kendall Nygard

Dr. Simone Ludwig

Dr. Peter G. Oduor

Approved:

23 August 2019
Date

Dr. Kendall Nygard
Department Chair

ABSTRACT

The resolution of spatial data has increased over the past decade making them more accurate in depicting landform features. From using a 60m resolution Landsat imagery to resolution close to a meter provided by data from Unmanned Aerial Systems, the number of pixels per area has increased drastically. Topographic features derived from high resolution remote sensing is relevant to measuring agricultural yield. However, conventional algorithms in Geographic Information Systems (GIS) used for processing digital elevation models (DEM) have severe limitations. Typically, 3-by-3 window sizes are used for evaluating the slope, aspect and curvature. Since this window size is very small compared to the resolution of the DEM, they are mostly resampled to a lower resolution to match the size of typical topographic features and decrease processing overheads. This results in low accuracy and limits the predictive ability of any model using such DEM data. In this dissertation, the landform attributes were derived over multiple scales using the concept of sliding window-based aggregation. Using aggregates from previous iteration increases the efficiency from linear to logarithmic thereby addressing scalability issues. The usefulness of DEM-derived topographic features within Random Forest models that predict agricultural yield was examined. The model utilized these derived topographic features and achieved the highest accuracy of 95.31% in predicting Normalized Difference Vegetation Index (NDVI) compared to a 51.89% for window size 3-by-3 in the conventional method. The efficacy of partial dependence plots (PDP) in terms of interpretability was also assessed. This aggregation methodology could serve as a suitable replacement for conventional landform evaluation techniques which mostly rely on reducing the DEM data to a lower resolution prior to data processing.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr.Anne Denton whose constant support and expert advise played a major role in guiding my research. Her enthusiasm to address the computational aspects in GIS is laudable and it played a vital role in making my doctoral research a success. I would also like to thank Dr. Peter Oduor and Dr. Stephanie Day. The knowledge I gained in Advanced GIS and Introduction to GIS classes were very useful in every step of my research. I am grateful to Dr. Kendall Nygard, Dr. Simone Ludwig and Dr. Peter Oduor for their willingness to serve as my graduate committee members, providing feedback, encouragement and constructive criticism.

I am grateful to NDEPSCoR and NSF for funding this research through award OIA-1355466. I would also like to thank NDSU Graduate School and College of Science and Mathematics for providing travel grants to the conferences I attended to present my research.

I would like to thank everyone in the Computer Science Department at NDSU for their guidance, support and wisdom. Special thanks to Karanam Ravichandran Dayananda; our technical discussions helped me hone my scientific research skills during my early days as a PhD student.

I am forever thankful to my teachers for their support, guidance and inspiration throughout my primary, middle and high school days at St. Xavier's Collegiate School.

Finally, I extend my gratitude to my parents and my wife Papia for their constant love, support, encouragement and sacrifices that made my dream come true.

DEDICATION

To my wife for introducing me to the world of GIS and her constant support and encouragement. My mother for her dedication and sacrifice to ensure that I understand the importance of education. My father for always believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX FIGURES

Figure Page

1. INTRODUCTION

1.1. Geographic Information Systems (GIS) and Remote Sensing

GIS is a way to gather, manage and analyze spatial data [2]. Google maps is an application utilizing spatial data. To travel from Point A to Point B, we enter the source and destination in the application. The application returns the shortest route between the selected locations with a few alternate routes as potential solutions. It uses spatial/geographical data, combines them with routing algorithms and outputs the shortest route. This is one of the countless benefits GIS has to offer. GIS is also used in agricultural yield prediction. We can use GIS tools and remotely sensed data to monitor the growth of crops and use the information to deduce if the yield would be low or high.

Two data formats that are available in GIS are vector data and raster data. Vector data is composed of points, lines and polygons [3]. The raster data is obtained and analyzed using remote sensing techniques. Remotely sensed data for an area is any data that is captured from a distance usually with sensors placed on-board satellites [4] or Lidar [5] from an aircraft. Two of the most commonly used raster data is a multispectral image and a Digital Elevation Model (DEM) [6]. Multispectral images are captured using sensors with several filters [7, 8]. These filters take images of the same study area only within a specific wavelength of light in the electromagnetic spectrum [9]. As surfaces vary in the wavelength of light they reflect back into the atmosphere, a multispectral image can be used for classifying land cover images very accurately [10]. They are used in variety of applications such as estimating forest canopy [11] and monitoring land-use change [12, 13]. Multispectral images can also be used to differentiate healthy and poor vegetation [14, 15]. An important index derived from a multispectral image and used in this dissertation is

Normalized Difference Vegetation Index (NDVI). This index is obtained using images captured in the Red and Infrared bands of a multispectral image and is useful to determine crop health [15].

1.2. Digital Elevation Model (DEM) derived attributes

DEMs are rasterized representations of a land surface elevation [16]. In a DEM raster, each pixel in the image represents a certain elevation value on the ground. A DEM can be used to derive certain landform attributes slope [17], aspect, [18] and curvature [19] which play a significant role in crop health analysis.

Slope of a land is defined as the maximum rate of change of elevation with respect to the surrounding. A slope can usually be positive, negative or zero. Slope in GIS however, is calculated to reflect the maximum change in angle of a location from its surrounding cell values. As such it can range between 0° and 90° where 0° denotes a flat surface [20]. The aspect is referred to as the compass direction that the slope [21] is facing. The curvature derived from a DEM is used to explain the rate of change of slope [22]. Usually this shows the type of slope the land is, convex or concave [23]. In GIS, these three attributes are calculated by running a 3-by-3 window across the entire DEM horizontally and applying the necessary equation for evaluation [24]. This means that each sub-window being evaluated has nine cells whose values are aggregated to come to a conclusion.

Each of these attributes derived from a DEM is important in estimating yield. Aspect can affect the duration of sunlight on a region thereby affecting the rate of evapotranspiration and significantly impacting crop yield [25]. A concave curvature may result in more water retention or snow depth than a convex one [26]. DEM values are used to derive the depth of the water table [27, 28], which may affect the frequency of irrigation [29]. Research indicates that plants thrive better below a certain slope angle [30] due to factors such as surface run-off [31]. GIS software

uses certain tool which are applied to a DEM to evaluate these attributes. Since these attributes are so important in crop yield analysis [32, 33], it is essential that the DEM and the GIS tools being used are reliable and accurate.

1.3. Drawbacks of existing GIS tools

The satellites used to capture these multispectral and DEM spatial data have come a long way. The Earth observing Multispectral Scanners (MSS) on-board Landsat satellites [34] launched in 1972 could capture very low resolution multispectral images close to 60m [35, 36]. This signifies that each pixel in the image resembled 60m on the ground. With the help of Lidar that can map data using a Nominal Pulse Spacing (NPS) of about 1m [37] we now have images that are several times higher in resolution than before. Hence a lot of information and data is produced for a similar study area. The problem, is that to process this high amount of data, GIS tools implement the same algorithms which are based on 3-by-3 sliding window analysis and developed decades ago. This introduces three issues.

As the resolution of DEMs and multispectral images keeps increasing, existing GIS tools that work on a 3-by-3 sliding window size could pick up a lot of noise in the data such as buildings and cars more accurately [38] compared to low resolution images used earlier since the average length of cars and some buildings can be smaller to show significant reflectance on a 80m resolution data. The second issue being that it takes a lot of time to process this information [39] since an area of 1 sq. km. on the ground has several times more information than before. The third issue arises from resampling the DEM. Resampling to a lower resolution is a common approach used by researchers to tackle high resolution spatial data. To reduce processing times, resolution of images are decreased. Without prior knowledge of which resolution is important for the study area, this

process can produce results with a lower accuracy [18] mostly because the resolution used for one study area may not be suitable for another based on its extent and topography.

This dissertation addresses these three issues by developing algorithms used to process remotely sensed data. A multi-scalar sliding window-based aggregation approach is proposed. This aggregation method is used to derive the GIS attributes slope, aspect and curvature from a DEM. The algorithm for deriving sliding window aggregates results in output on every scale, thereby enabling multi-scalar approaches that makes use of the intermediate results. The output obtained from several scales are used to study their impact on crop productivity and the results are compared with the traditional method.

1.4. Multiscalar sliding window-based aggregation

Sliding window analysis is a technique used in image processing [40]. It works by evaluating a group of pixels using a set window size. A 3-by-3 window would evaluate 9 cells in the upper left corner of the image and slide horizontally to the right, one column at a time performing the same evaluation. The number of steps is also referred to as the stride. A stride of one is commonly used but various machine learning models such as Convolutional Neural Networks allows the user to modify the stride [41] to discover more patterns. Using this sliding window we can run various filters to find certain patterns in the image. For example, a max filter would return the maximum values across the entire image when evaluated using a 3-by-3 sliding window.

In this research, the multi-scalar aspect of sliding window analysis focuses on reusing results from previous iteration. An example of the multi-scalar sliding window-based aggregation is visualized in Figure 1.1. Result obtained from a 2-by-2 sliding window was used to evaluate a 4-by-4 window. Again, the output from window size 4-by-4 was used for evaluating the window size 8-by-8 and so on. This method of reusing aggregates is possible for evaluating means, and

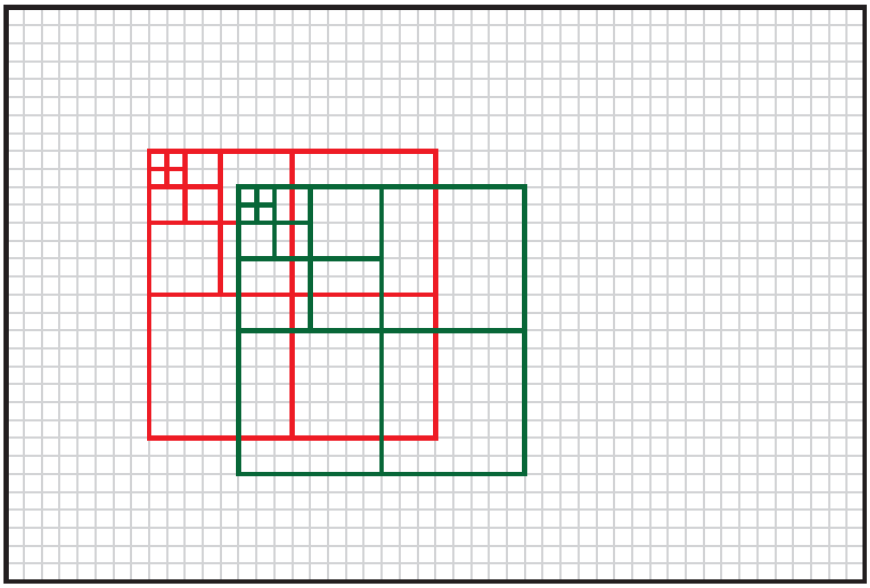


Figure 1.1. Sliding window aggregation shows how results from a 2-by-2 grid are used as input for a 4-by-4 grid. Likewise the results of 4-by-4 are reused in 8-by-8 evaluation.

does not work for medians. Since the GIS tools also utilize mean, the idea of reusing aggregates can be extended to evaluate slope, aspect and curvature.

1.5. Proposed approach

In this dissertation, the multi-scalar sliding window-based aggregation has been extended to evaluate GIS attributes slope, curvature and aspect. The objective for developing the algorithm was to ensure that the landform attributes being evaluated can scale well to high resolution datasets. The NDVI was also aggregated for the similar windows.

Machine learning models such as Random Forest [42] and Naive Bayes [43] were used on the derived slope, aspect and curvature output from the sliding window-based aggregation. These models were used to perform predictive analysis by studying their impact on NDVI. The objective of predictive analysis was to observe how effective the results derived from multi-scalar sliding window-based aggregation were in predicting yield compared to the conventional approach. The output from predictive analysis was then visualized using partial dependence plots (PDP) [44] to

show how the individual attributes effect NDVI. Finally, these prediction models were compared to similar models derived from results obtained using the traditional approach in ArcGIS. This comparison was done to evaluate if the proposed model was better than the conventional one.

1.6. Overall contribution

To evaluate the relationship of yield with the landform attributes derived from a DEM, several classification models [45, 46, 47] with heuristics have been used earlier. DEM and NDVI data is the foundation to derive these models that aid in a farmers' decision-making process [16]. Due to time constraints the resampling images to a lower resolution has become a very common trend. Using the proposed approach, researchers can now obtain DEM or multispectral images over several scales which offers three advantages. Firstly, it gives any analyst the freedom to choose the most appropriate scale for their study without sacrificing resolution. Secondly, since the process reuses values obtained from a previous iteration to obtain multiple outputs it achieves a logarithmic run-time efficiency. Finally, integrating the algorithm in commercially available GIS software would allow multiple outputs and faster processing which was not possible earlier. The benefits of this method is not limited to agriculture. Any research that utilizes raster data can use the multi-scalar analysis. GIS researchers and scholars would be able to make a logical choice using this algorithm compared to the conventional approach of selecting a lower resolution image.

1.7. Organization of the dissertation

This dissertation is composed of six chapters. The first chapter contains the general introduction. The next four chapters are related to three published papers [48, 49, 50] and two papers which have been submitted for review. The last chapter contains a discussion of general conclusions and future work.

Chapter one contained the general introduction of what is GIS and Remote Sensing and also explained the difference between multispectral images and DEM data that was used in this research. A brief introduction of sliding window analysis was presented with an explanation of how it can be extended to perform a multi-scalar sliding window-based aggregation. The introduction also discussed the benefits of this research and talked briefly about the proposed approach.

Chapter two discussed how the landform attributes slope, aspect and curvature were derived using the multi-scalar sliding window-based aggregation. Three of the most common curvature types were derived. They were profile curvature, planform curvature and mean curvature. The process began by aggregating four values in each iteration, and aggregates from previous iterations were reused as shown in Figure 1.1. The proposed method utilized window scales of $w = 4, 8, 16, 32$ and 64 to derive the landform attributes. The derived outputs were compared with the output from ArcGIS in comparable scales of $w = 3, 9, 15, 30$ and 63 to establish the benefit of the proposed methodology. The usefulness of the proposed strategy was demonstrated on Digital Elevation Model data.

After deriving the landform attributes slope, aspect and curvature; the next step was to perform an error analysis on the derived results. Like any raster data, a DEM also suffers from errors and these errors can propagate on changing window sizes. In Chapter three, the propagation of error in the DEM arising due to several length scales while increasing the window size was investigated. Root Mean Square Error (RMSE) propagation was recorded during this process. To establish how much information the proposed method can retain, semivariograms were derived from the DEM data and compared with the results obtained from the traditional method to justify the importance of multi-scalar sliding window-based aggregation.

In chapter four, the effectiveness of the aggregation method in generating predictive models was investigated. Random Forest and Naive Bayes classification were implemented from the DEM derived attributes slope, aspect and curvature and used to predict NDVI. Results from this evaluation were compared with the predictive models derived from the proposed method to ascertain how reliable the multi-scalar datasets were and if the proposed approach was a suitable alternative to the traditional approach used in several GIS software.

In chapter five, the consistency of multi-scalar aggregation was demonstrated by generating predictive models on datasets from several study areas with varying extent. The results of the predictive modelling was also visualized using partial dependence plots to establish which GIS attributes played a significant role in yield prediction.

Finally, in chapter six, the findings from this dissertation were summarized and the proposed future work which can be done to extend this research was discussed.

2. DERIVING TOPOGRAPHIC VARIABLES USING SLIDING WINDOW-BASED AGGREGATION

2.1. Introduction

Slope, aspect and curvature are important variables used extensively to explain various landform features. As the length scale used for analysis of a DEM can produce significant variation in topographical estimation and error propagation [51], using the concept of scaling results for curvature, aspect, and slope estimation seems highly justifiable. Scaling these raster datasets across multiple window sizes simultaneously gives researchers the freedom to choose the most informative scale for analysis without losing any features from the DEM. This chapter summarizes the algorithm used to calculate the landform attributes slope, aspect and curvature on a multi-scalar level.

2.1.1. Slope

A DEM is used for creating various features such as slope, rate of water flow, etc. [52]. Slope evaluation has been regarded as an integral factor for works related to watershed delineation [53]. In Revised Universal Soil Loss Equation (RUSLE), erosion was represented as an exponential function of the slope [54]. This made erosion and soil loss highly dependent on slope. The authors in [55] reported that an increase in slope estimation error of 10% can increase the soil-loss estimates error by as much as 20%.

A study on two-slope calculation was conducted to evaluate the results produced by different slope algorithms on Lidar-derived DEM data [56] using DEM resolution as a factor. The interpolation techniques used to generate 1m, 5m, and 10m Lidar-derived DEMs and the 1m, 10 m aerial DEMs can affect the outcome [57]. Prevalent interpolation techniques such as nearest

neighbor, cubic convolution and bilinear interpolation [58] are commonly used in GIS software. It is imperative to understand which interpolation has to be applied for the task as they can produce different outcome [59]. One interpolation technique can be better in generating a smoother surface than another [60]. Certain interpolation techniques are more sensitive along the DEM borders and can generate inconsistent results compared to the inner regions in the DEM [61]. Since resampling DEMs to a lower resolution can adversely affect results, maintaining the integrity of DEMs and using separate window scales for slope evaluation can be a suitable alternative.

Average estimation and maximum difference in elevation are frequently used in slope evaluation [62]. Averaging slope methods usually utilize all the pixels in the window size to calculate the slope for the middle cell. Maximum slope methodology calculates slope by comparing the central cell with the other cell values in a certain window size that shows the maximum difference in elevation [63, 24]. While performing slope evaluation, the scale also plays a significant role. Using a 3-by-3 fixed window to obtain slope, on a high-resolution DEM derived from Lidar point cloud can increase computation time and generate results which might not be relevant to the study. Since slope is dependent on the elevation data [54, 64] a 3-by-3 window on high resolution DEMs from Lidar is also affected by noise as the dataset is sensitive to man-made features, artificial ridges and grooves. In [56] the authors established a direct correlation between the DEM resolutions, type of slope used and the variation in results. Averaging Neighborhood Slope (ANS) methodology using a 3-by-3 window produced erroneous results that showed high difference in the actual and calculated slope [65]. Since the 3-by-3 moving window ignores elevation value in the middle cell, abnormally high slope values were reported near streams as compared to slopes that were reported in smooth areas with low difference in elevation between adjacent cells [65]. This problem had

also been reported in calculation of slopes in hilly regions where there is a sudden drop in the elevation such as ridges and places of sudden rise in elevation such as peaks [62].

2.1.2. Aspect

In ArcGIS aspect is evaluated on a raster data by finding the downslope direction of the maximum rate of change of a cell value compared to its neighbors [21]. Like the slope, the aspect is an important factor used to determine the growth of crops [66] as it has an effect on the sunlight striking on a surface. Research conducted in the Qilian Mountain area in China established that regions between a certain extent of North-East and North-West showed a higher vegetation growth compared to other regions [25]. Aspect also has an affect on the local temperature as hills with westerly aspect are warmer compared to slopes facing eastward. The effect of aspect is very much predominant in regions like the Himalayas where southerly slopes have higher vegetation since they are shielded from the cold dry winds of the North. Research in [67] established that grass on slopes with southerly aspect on British chalk grasslands are more resilient to extinction due to warm and drought conditions. Forests are readily present in easterly aspect in Australia as they are not facing the dry winds approaching from the West [68].

If there is a high variation in local weather based on dry winds or sunlight, a farmer can make a logical choice to plant crops in regions where the aspect does not face hot and dry winds to see better productivity. Aspect and slope, combined with precipitation can decide surface run-off which is an important factor to determine soil erosion [69]. ArcGIS highlights eight aspect options based on the direction the slope faces. They are Flat, North, North-East, East, South-East, South, South-West, West, and North-West. Like the slope, it uses a fixed 3-by-3 sub-cell window for evaluation.

2.1.3. Curvature

Curvature defined as the rate of change of slope is calculated from the raster DEMs where pixel values correspond to land elevation [70]. However, there is not just a single curvature that can be calculated. In [23], the author mentions up to nine curvature types such as profile, plan, maximum, minimum, longitudinal, medial, general, transverse, and tangential that can be evaluated. Profile curvature refers to variation of slope on a vertical plane [71, 72] whereas plan curvature refers on a horizontal plane [73]. An important observation is that the result obtained from curvature applications widely vary from one method to another. Commercial GIS software such as ArcGIS and Jenness refer to a concave and convex curvature as positive and negative respectively. On the other hand, Landserf and SAGA systems use the opposite sign convention. This is a huge problem across multiple software systems that use different curvature algorithms making interoperability difficult. Tangential curvature [74] is mostly like plan curvature, but they are highly suited for horizontal flow analysis. Longitudinal curvature has similar geomorphologic resemblance with the profile curvature [23]. General, maximum, and minimum curvatures as their names suggest are calculated by running a 3-by-3 fixed sliding window on the elevation data. These curvatures are used by Landserf, and implemented by Wood [75]. The general curvature [76] combines both profile and plan curvature for a generalized result and is widely used for curvature evaluation. Even though several curvature systems exist, the consistent approach made by all these algorithms includes using a fixed sliding window size preferably a 3-by-3 cell window for calculation. The two most common curvature equations are those of Evan's and Florinsky's [77, 78] that uses six polynomial parameters and Zevenbergen's [79] utilizing nine parameters. The latter is utilized by the curvature tool in ArcMap. Both approaches are also restricted to a 3-by-3 cell window size.

2.2. Previous work

There is compelling evidence to support the application of multi-scalar analysis in the field of GIS and remote sensing [80]. Since GIS algorithms for evaluating curvature and slope use a fixed window size, experiments with variable window sizes and their impact on results were also conducted [75, 51]. In a study conducted by Wood [75], window sizes in powers of two were used to perform terrain analysis. The results generated a function to filter out high frequency noise in the dataset. This idea builds on even window sizes and makes it easier to explore the impact of having a higher windows on the output. Using several large windows, Wood [75] was able to generalize the DEM and obtain a macroscopic view which expressed patterns otherwise obscured by errors (sinks) in the DEM.

Since the GIS attributes can be calculated by fitting a least squares equation to the DEM window, an error propagation model was developed based on Taylor approximation of least squares fitting in [51]. Monte Carlo simulation was used to derive the error between the points obtained from interpolation and the actual vector points fed to create the DEMs across variable window sizes. The authors in [51] observed a high degree of correlation between elevation value with the window sizes and concluded that errors in the DEMs have much less impact when a window size greater than seven is used for study.

The concept of sliding window analysis implemented in [1] was extended in this research to derive GIS attributes slope, aspect and curvature. The sliding window concept was used in [1] as an aggregation technique for the computation of regression and correlation lines across multiple window scales. The correlation and regression lines were derived for Red vs Near-Infrared (NIR) bands of a multispectral image followed by another analysis between yield and NDVI. As the window size used for the experiment doubled in each iteration, the aggregates obtained from

the previous iteration were used for the next window size, making the algorithms run-time efficiency logarithmic. This method was tested for efficiency against the conventional technique using GRASS [81]. A DEM of 1024-by-1024 pixels was evaluated for window sizes $4, 8, 16, 32$ and 64 respectively. Application of the algorithm without using linear aggregates from the previous iteration made it scale linearly as shown in Figure 2.1 compared to the logarithmic scaling of sliding window-based aggregation.

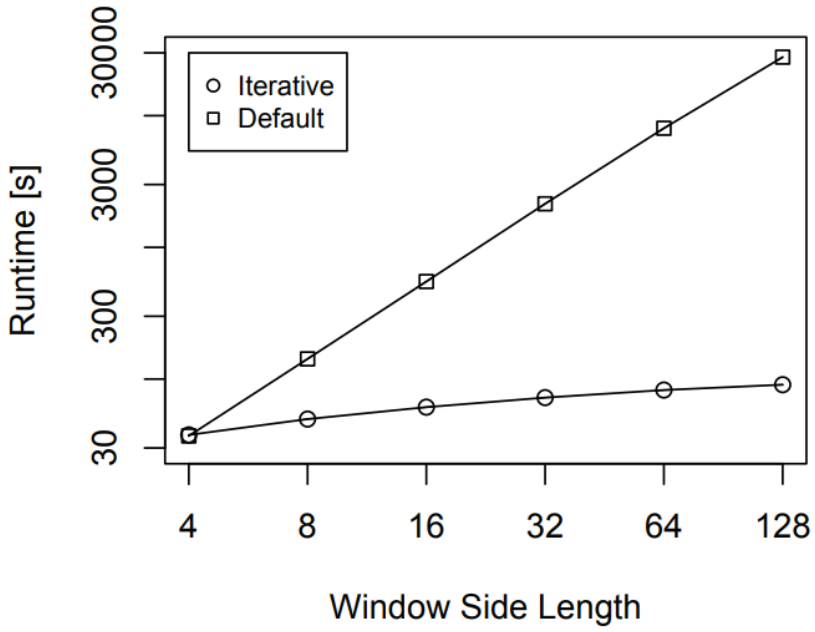


Figure 2.1. Multi-scalar aggregation from [1] shows logarithmic performance efficiency when compared with traditional method.

2.3. Deriving landform attributes

2.3.1. Slope

To evaluate the slope of regression, a least squares fit was evaluated for z-values with respect to both x and y-axis as shown in equation. 2.1 by minimizing the squared error [49].

$$z_{lin}(x,y) = \begin{pmatrix} b_0 & b_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + c_s \tag{2.1}$$

Here, b_0 represents slope for x as independent variable while b_1 represents the slope for y as an independent variable. The minimization was performed by taking partial derivatives with respect to the 3 parameters $b_0, b_1,$, and c and setting them to zero. To solve for b_0 equation 2.2 was evaluated.

$$\frac{\partial S}{\partial b_0} = \sum (z - b_0 x - b_1 y - c_s) \times (-2x) = 0$$

$$\sum zx - b_0 \sum x^2 - b_1 \sum xy - c_s \sum x = 0 \tag{2.2}$$

To solve for b_1 we evaluated equation 2.3.

$$\frac{\partial S}{\partial b_1} = \sum (z - b_0 x - b_1 y - c_s) \times (-2y) = 0$$

$$\sum zy - b_0 \sum xy - b_1 \sum y^2 - c_s \sum y = 0 \tag{2.3}$$

Similarly for c_s equation 2.4 was evaluated.

$$\frac{\partial S}{\partial c_s} = \sum (z - b_0 x - b_1 y - c_s) \times (-2) = 0$$

$$\sum z - b_0 \sum x - b_1 \sum y - c_s = 0 \tag{2.4}$$

Solving these partial derivatives required aggregating several $\sum x$ or $\sum y$ values in a sub-cell window. Consider a 4-by-4 sub-window as shown in Figure 2.2. Z_i represents the elevation value of each cell. Since the coordinates of x and y are symmetrical about the center, the summation of all the x and y values for the 16 cells was zero. Every sub-window in the sliding window-based algorithm was evaluated using this scheme. Hence, any terms containing $\sum x$ or $\sum y$ would evaluate to zero.

Using this symmetry and rewriting equation 2.2 gave

$$\sum zx - b_0 \sum x^2 = 0 \tag{2.5}$$

Rewriting equation 2.3 gave

$$\sum zy - b_1 \sum y^2 = 0 \tag{2.6}$$

Finally rewriting equation 2.4 gave

$$\sum z - c_s = 0 \tag{2.7}$$

Removing the $\sum x$ and $\sum y$ terms during the evaluation of these partial derivatives yields the following results as shown in equation 2.8

$$\begin{aligned} b_0 &= \frac{\sum xz}{\sum x^2} \\ b_1 &= \frac{\sum yz}{\sum y^2} \\ c_s &= \sum z \end{aligned} \tag{2.8}$$

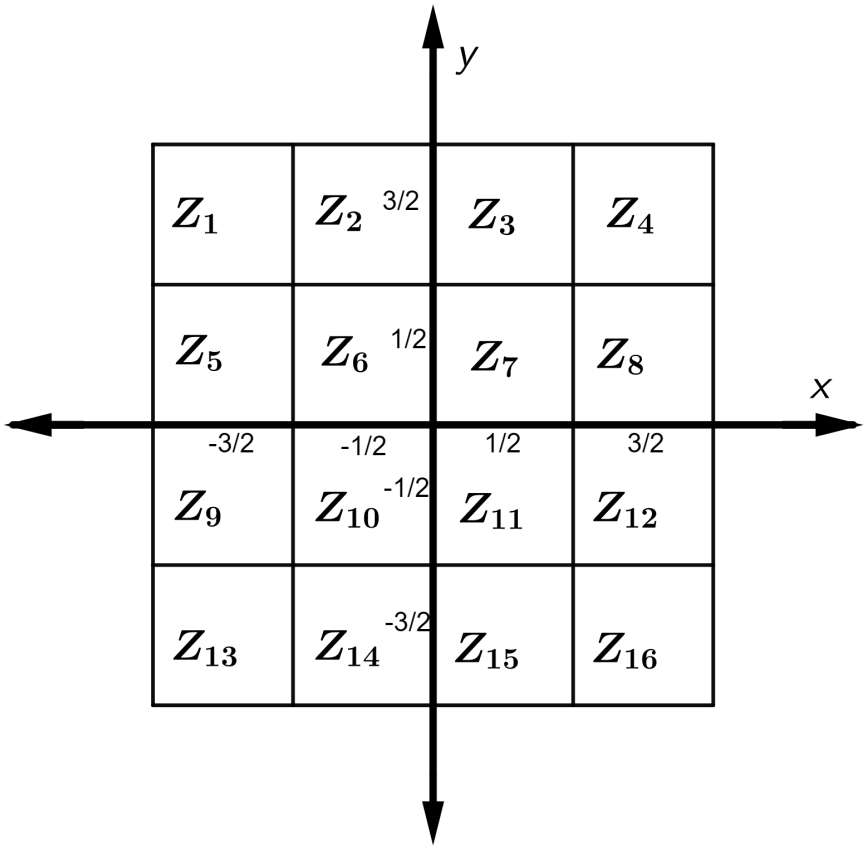


Figure 2.2. Window aggregation scheme for $\sum x$ and $\sum y$ values.

Equation 2.9 shows the slope of a line along two dimensions.

$$\begin{aligned} slope &= \arctan(b_0 \frac{\sum(xz)}{\sqrt{\sum x^2 + \sum y^2}} + b_1 \frac{\sum(yz)}{\sqrt{\sum x^2 + \sum y^2}}) \\ &= \arctan\left(\frac{b_0 \sum xz + b_1 \sum yz}{\sqrt{\sum x^2 + \sum y^2}}\right) \end{aligned} \tag{2.9}$$

Since the values for b_0 and b_1 were evaluated earlier and $\sum x^2$, $\sum y^2$ represents the same output due to symmetry, equation 2.9 was further simplified to equation 2.10. This equation was used for the evaluation of the slope of regression.

$$slope = \arctan\left(\frac{\sqrt{\sum(xz)^2 + \sum(yz)^2}}{\sum x^2}\right) \tag{2.10}$$

The simplified equation 2.10 for calculating slope produced three new terms. The numerator has $\sum(xz)^2$ which is the change in z values along x-axis while increasing the length-scale in terms of 2^i where $i = 2,3,4,5$ and 6. $\sum yz$ used the same concept of increasing the length-scale but showed the change of z values along y-axis. These terms had to be aggregated to accommodate the coordinate shift and ensure that the origin remained fixed in the sub-window being evaluated.

Figure 2.3 shows an example of two 4-by-4 windows (top row) which are aggregated to yield an 8-by-8 window (bottom row). The 4-by-4 windows are two positive y-axis quadrants for the newly derived 8-by-8 window. The axis $y = 0$ and $x = 0$ for each of the small sub-windows needs to be shifted to ensure the coordinate system aligns properly for the larger window scale. The quadrants containing z-values for the window scale being evaluated was represented with respect to x and y using equation 2.11.

$$z(x,y) = \begin{cases} z_{00}(x_0,y_0) \text{ for } x < 0, y < 0 \\ z_{10}(x_1,y_0) \text{ for } x > 0, y < 0 \\ z_{01}(x_0,y_1) \text{ for } x < 0, y > 0 \\ z_{11}(x_1,y_1) \text{ for } x > 0, y > 0 \end{cases} \tag{2.11}$$

The summation of results obtained from quadrants in the window scale 4 was represented as (x_0,y_0) bottom-left, (x_1,y_0) bottom-right, (x_0,y_1) top-left, and (x_1,y_1) top-right. Quadrants

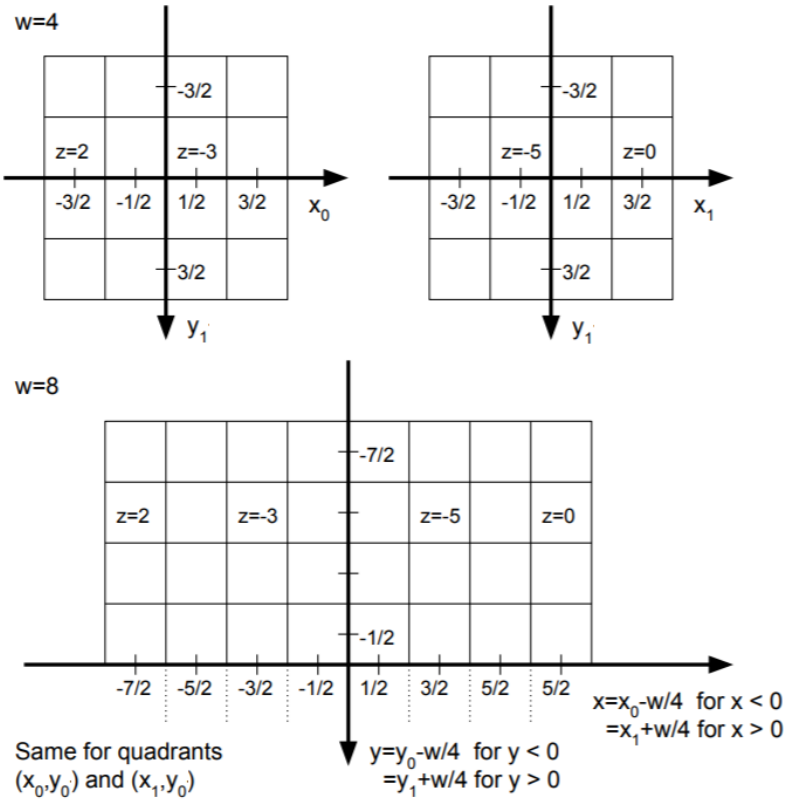


Figure 2.3. Window aggregation scheme for upscaling. Results of two 4-by-4 windows (top) are combined to show their coordinate shift in the 8-by-8 window (bottom).

(x_0, y_1) top-left, and (x_1, y_1) top-right are shown in Figure 2.3 (top-row). The terms on the right side of the equation are not single values, rather they were derived from the summation of xz or yz performed in the previous length scale of size $\frac{w}{2}$ which is 4 in this example. The shift to transform axis was achieved by the relations shown in equation 2.12. To further explain this method in action, let us consider the quadrant of the 8-by-8 window denoted by x_0, y_1 (top-left). All xz terms in this quadrant needs to be shifted by $-\frac{w}{4}$ and all yz terms should be shifted by $+\frac{w}{4}$ to compensate for the increase in window scale from 4 to 8. The process was repeated for the remaining three quadrants but with the proper sign convention as shown in equation 2.12. Combining results of xz and yz from the 4 quadrants of window scale $\frac{w}{2}$ the xz or yz for the present window scale w could be evaluated as shown in equation 2.13 and equation 2.14.

$$
\begin{aligned}
x &= x_0 - \frac{w}{4} \quad for \quad x < 0 \\
x &= x_1 + \frac{w}{4} \quad for \quad x > 0 \\
y &= y_1 + \frac{w}{4} \quad for \quad y > 0 \\
y &= y_0 - \frac{w}{4} \quad for \quad y < 0
\end{aligned} \tag{2.12}
$$

$$
\begin{aligned}
\sum xz = \frac{1}{4}\Big(\sum(\Big(x_0 - \frac{w}{4}\Big) z_{00} \; + \sum \Big(x_1 + \frac{w}{4}\Big) z_{10} \\
+ \sum \Big(x_0 - \frac{w}{4}\Big) z_{01} + \sum \Big(x_1 + \frac{w}{4}\Big) z_{11}\Big)
\end{aligned} \tag{2.13}
$$

$$
\begin{aligned}
\sum yz = \frac{1}{4}\Big(\sum \Big(y_0 - \frac{w}{4}\Big) z_{00} \; + \sum \Big(y_0 - \frac{w}{4}\Big) z_{10} \\
+ \sum \Big(y_1 + \frac{w}{4}\Big) z_{01} + \sum \Big(y_1 + \frac{w}{4}\Big) z_{11}\Big)
\end{aligned} \tag{2.14}
$$

To evaluate $\sum x^2$ in the denominator of equation 2.10, equation 2.15 was used. Applying the formula on Figure 2.2 with $w = 4$, gives 20 as the result. In this equation, $k - \frac{1}{2}$ was evaluated instead of k because the centroid coordinates of z-values are not whole numbers. For example, the coordinates of Z_7 are $(0.5, 0.5)$ for k = 1. The term $\frac{1}{w^2}$ outside the summation is used to normalize the result across multiple windows as the total cells being used for analysis is square of the window size.

$$
\sum x^2 = \sum y^2 = \frac{1}{w^2} 2w \sum_{k=1}^{\frac{w}{2}} (k - \frac{1}{2})^2 \tag{2.15}
$$

Since a stride of one was used to evaluate sub-windows and re-center the origin every time, equation 2.15 could be generalized further and used across all sub-windows as the evaluation was symmetrical like $\sum x$ and $\sum y$. This summation term in equation 2.15 can be computed based on power sums [82] as shown in equation 2.16 [83]. The result of this evaluation is shown in equation 2.17. The $\sum x^2$ was constant across all the sub-window calculations.

$$\sum_{k=1}^{n} k = \frac{1}{2}\left(n^2+n\right)$$
$$\sum_{k=1}^{n} k^2 = \frac{1}{6}\left(2n^3+3n^2+n\right) \tag{2.16}$$

$$\sum x^2 = \sum y^2 = \frac{2}{w}\left(\sum_{k=1}^{\frac{w}{2}} k^2 - \sum_{k=1}^{\frac{w}{2}} k + \frac{w}{8}\right) = \frac{w^2-1}{12} \tag{2.17}$$

2.3.2. Aspect

Aspect (α) is the compass direction of slope which specifies the direction in which the slope is facing. Aspect finds many applications in agriculture as the slope direction might have a significant impact on vegetation or irrigation required to sustain a good yield [25]. Aspect was evaluated along four directions, North-East (NE), North-West (NW), South-East (SE) and South-West (SW) [49].

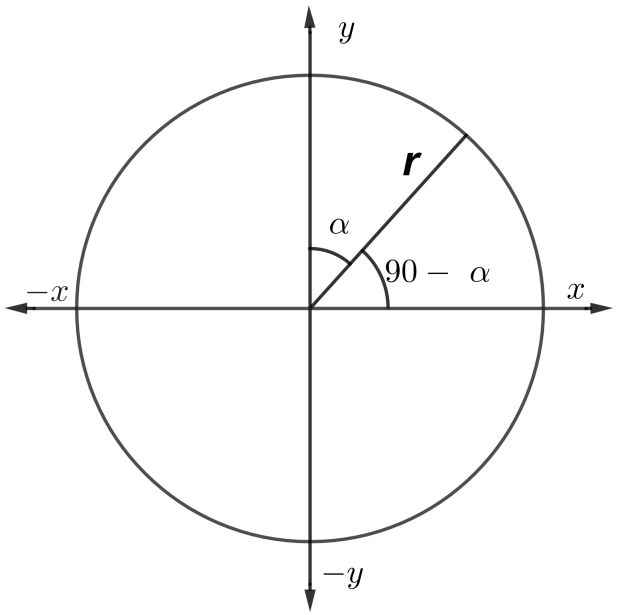


Figure 2.4. Aspect α is evaluated counter-clockwise.

The usual equation for x and y directions based on x axis is given by equation 2.18. The terms are negative due to the clockwise direction as shown in Figure 2.4.

$$x = \mathbf{r}\cos[-(90-\alpha)]$$

$$y = \mathbf{r}\sin[-(90-\alpha)] \tag{2.18}$$

Considering r as constant, $\cos-(\theta) = \cos(\theta)$ and $\sin-(\theta) = -\sin(\theta)$ we can rewrite equation 2.18 as $x = \sin(\alpha)$ and $y = -\cos(\alpha)$.

Substituting these new values of x, y and b_0, b_1 obtained earlier in equation 2.8 into equation 2.1 we derive equation 2.19. Differentiating both sides with respect to α we get and substituting values from equation 2.8 gave $\tan(\alpha) = -\frac{b_0}{b_1}$.

$$z_{lin}(x,y) = \begin{pmatrix} \frac{\sum xz}{\sum x^2} & \frac{\sum yz}{\sum y^2} \end{pmatrix} \begin{pmatrix} \sin(\alpha) \\ -\cos(\alpha) \end{pmatrix} + \sum z$$

$$z_{lin}(x,y) = \frac{\sum xz}{\sum x^2}\sin(\alpha) - \frac{\sum yz}{\sum y^2}\cos(\alpha) + \sum z \tag{2.19}$$

Equation 2.20 shows the final aspect values after incorporating the directions.

$$\alpha = \begin{cases} \pi - \arctan\frac{\sum xz}{\sum yz} \; \forall \sum yz < 0 \; (SW \;\&\; SE) \\ 2\pi - \arctan\frac{\sum xz}{\sum yz} \; \forall \sum xz > 0, \; \sum yz > 0 \; (NW) \\ -\arctan\frac{\sum xz}{\sum yz} \; \forall \sum xz < 0, \; \sum yz > 0 \; (NE) \end{cases} \tag{2.20}$$

2.3.3. Curvature

To evaluate the profile, plan and mean curvatures, a similar process of least squares fit was performed. However instead of solving a linear problem; a quadratic one was solved as shown in equation 2.21. This was followed by minimizing squared error as before but now there were more constants as compared to slope evaluation.

$$z_{quad}(x,y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a_{00} & a_{10} \\ a_{10} & a_{11} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_{c0} & b_{c1} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + c_c \tag{2.21}$$

Partial derivatives were evaluated with respect to a_{00} as shown in equation 2.22. Any term with odd powers of x and y are zero due to symmetry as shown earlier in Figure 2.2. Hence, terms like $\sum x^3$ were zero. Similarly, any term having both x and y where either of them had an odd power would produce zero due to symmetry. So, terms such as $\sum x^3 y$ and $\sum x^2 y$ were also zero. Partial derivatives with respect to a_{00} is shown in equation 2.22. Partial derivatives with respect to a_{10} is shown in equation 2.23. As observed during the slope evaluation, some of the terms reduced to zero due to symmetry.

$$\begin{aligned} \frac{\partial C}{\partial a_{00}} = \sum -2x^2(z - a_{00}x^2 - 2a_{10}xy &- a_{11}y^2 - b_{c0}x - b_{c1}y - c_c) = 0 \\ \sum zx^2 - \sum a_{00}x^4 - 0 - \sum a_{11}x^2y^2 &- 0 - 0 - \sum c_c x^2 = 0 \\ \sum zx^2 = \sum a_{00}x^4 + \sum a_{11}x^2y^2 &+ \sum c_c x^2 \end{aligned} \tag{2.22}$$

$$\begin{aligned} \frac{\partial C}{\partial a_{10}} = -4xy(\sum z &- \sum a_{00}x - \sum 2a_{10}xy - \sum a_{11}y^2 - \sum b_{c0}x - \sum b_{c1}y - \sum c_c) = 0 \\ \sum zxy &- 0 - \sum 2a_{10}x^2y^2 - 0 - 0 - 0 - 0 = 0 \\ a_{10} &= \frac{\sum xzy}{\sum 2x^2y^2} \end{aligned} \tag{2.23}$$

Partial derivatives with respect to a_{11}, b_{c0}, $b_{c1}$and b_{c_c} are shown in equation 2.24, equation 2.25, equation 2.26, and equation 2.27 respectively.

$$\begin{aligned}
\frac{\partial C}{\partial a_{11}} = -2y^2(\sum z &- \sum a_{00}x^2 - \sum 2a_{10}xy - \sum a_{11}y^2 - \sum b_{c0}x - \sum b_{c1}y - \sum c_c) = 0 \\
\sum zy^2 &- \sum a_{00}x^2y^2 - 0 - \sum a_{11}y^4 - 0 - 0 - \sum c_c y^2 = 0 \\
\sum zy^2 &= \sum a_{00}x^2y^2 + \sum a_{11}y^4 + \sum c_c y^2
\end{aligned} \tag{2.24}$$

$$\begin{aligned}
\frac{\partial C}{\partial b_{c0}} = -2x(\sum z &- \sum a_{00}x^2 - \sum 2a_{10}xy - \sum a_{11}y^2 - \sum b_{c0}x - \sum b_{c1}y - \sum c_c) = 0 \\
\sum xz &- 0 - 0 - 0 - \sum b_{c0}x^2 - 0 - 0 = 0 \\
b_{c0} &= \frac{\sum xz}{\sum x^2}
\end{aligned} \tag{2.25}$$

$$\begin{aligned}
\frac{\partial C}{\partial b_{c1}} = -2y(\sum z &- \sum a_{00}x^2 - \sum 2a_{10}xy - \sum a_{11}y^2 - \sum b_{c0}x - \sum b_{c1}y - \sum c_c) = 0 \\
\sum zy &- 0 - 0 - 0 - 0 - b_{c1}y^2 - 0 = 0 \\
b_{c1} &= \frac{\sum zy}{\sum y^2}
\end{aligned} \tag{2.26}$$

$$\begin{aligned}
\frac{\partial C}{\partial c_c} = -2(\sum z &- \sum a_{00}x^2 - \sum 2a_{10}xy - \sum a_{11}y^2 - \sum b_{c0}x - \sum b_{c1}y - 1) = 0 \\
\sum z &- \sum a_{00}x^2 - 0 - \sum a_{11}y^2 - 0 - 0 - 1 = 0 \\
\sum z &= \sum a_{00}x^2 + \sum a_{11}y^2 + 1
\end{aligned} \tag{2.27}$$

Multiplying equation 2.27 with y^2 and subtracting from equation 2.24 we get equation 2.28

$$a_{11} = \frac{\sum y^2 z - \sum x^2 \sum z}{\sum y^4 - \sum (x^2)^2} \tag{2.28}$$

Similarly a_{11} was evaluated by multiplying equation 2.27 by x^2 and subtracting from equation 2.23. The result is shown in equation 2.29

$$a_{00} = \frac{\sum x^2 z - \sum z \sum x^2}{\sum x^4 - \sum (x^2)^2} \tag{2.29}$$

Finally substituting the newly obtained values of a_{00} and a_{11} to equation 2.22 equation 2.30 was derived.

$$c_c = \sum z - \frac{\sum x^2 \sum x^2 z + \sum x^2 \sum y^2 z - 2(\sum x^2)^2 \sum z}{\sum x^4 - (\sum x^2)^2} \tag{2.30}$$

The profile curvature in the direction of the slope was evaluated using equation 2.31. Planform curvature in a direction perpendicular to the slope is shown using equation 2.32.

$$Curv_{Profile} = \frac{a_{00}(\sum xz)^2 + 2a_{10} \sum xz \sum yz + a_{11}(\sum yz)^2}{(\sum xz)^2 + (\sum yz)^2} \tag{2.31}$$

$$Curv_{Plan} = \frac{a_{00}(\sum yz)^2 - 2a_{10} \sum xz \sum yz + a_{11}(\sum xz)^2}{(\sum xz)^2 + (\sum yz)^2} \tag{2.32}$$

The mean curvature obtained as an average for both and used for this analysis is shown in equation 2.33.

$$Curv_{mean} = \frac{\sum x^2 z + \sum y^2 z - 2 \sum x^2 \sum z}{\sum x^4 - (\sum x^2)^2} \tag{2.33}$$

The fourth order term x^4 shown in equation 2.34 was derived using power sums similar to the process used for deriving x^2 earlier in equation 2.16. Unlike equation 2.16 where power sums were evaluated over $\sum k$ and $\sum k^2$, this term also had two additional power sums $\sum k^3$ and $\sum k^4$ as shown in equation 2.34.

$$\sum x^4 = \frac{1}{w^2} 2w \sum_{k=1}^{\frac{w}{2}} (k - \frac{1}{2})^4 \tag{2.34}$$

According to power sums, $\sum k^3$ and $\sum k^4$ when expanded is shown in equation 2.35.

$$\begin{aligned}\sum_{k=1}^{n} k^3 &= \frac{1}{4}\sum(n^4+2n^3+n^2)\\ \sum_{k=1}^{n} k^4 &= \frac{1}{30}\sum(6n^5+15n^4+10n^3-n)\end{aligned} \tag{2.35}$$

Solving for $\sum x^4$ equation 2.36 was derived.

$$\sum x^4 = \frac{3w^4-10w^2+7}{240} \tag{2.36}$$

In $Curv_{mean}$, the denominator terms $\sum x^4 - (\sum x^2)^2$ were obtained using equation 2.36 and equation 2.17 and shown below. This term was zero for $w = 2$ which makes sense as curvature cannot be calculated on two windows.

$$\sum x^4 - (\sum x^2)^2 = \frac{w^4-5w^2+4}{180} \tag{2.37}$$

The terms $\sum x^2 z$, $\sum y^2 z$ and $\sum xyz$ used for curvature evaluation were also aggregated using window-based aggregation in the same manner as done for slope in equation 2.13 and equation 2.14. The result obtained is shown in equation 2.38. $\sum z$ was accumulated as shown in equation 2.39

$$\begin{aligned}\sum x^2 z &= \frac{1}{4}(\sum(x_0-\frac{w}{4})^2 z_{00}+\sum(x_1+\frac{w}{4})^2 z_{10}\\ &+ \sum(x_0-\frac{w}{4})^2 z_{01}+\sum(x_1+\frac{w}{4})^2 z_{11})\\ \sum y^2 z &= \frac{1}{4}(\sum(y_0-\frac{w}{4})^2 z_{00}+\sum(y_1-\frac{w}{4})^2 z_{10}\\ &+ \sum(y_0+\frac{w}{4})^2 z_{01}+\sum(y_1+\frac{w}{4})^2 z_{11})\\ \sum xyz &= \frac{1}{4}(\sum(x_0-\frac{w}{4})(y_0-\frac{w}{4}) z_{00}+\sum(x_1+\frac{w}{4})(y_0-\frac{w}{4}) z_{10}\\ &+ \sum(x_0-\frac{w}{4})(y_1+\frac{w}{4}) z_{01}+\sum(x_1+\frac{w}{4})(y_1+\frac{w}{4}) z_{11})\end{aligned} \tag{2.38}$$

$$\sum z = \frac{1}{4}\left(\sum z_{00}+\sum z_{10}+\sum z_{01}+\sum z_{11}\right) \tag{2.39}$$

2.4. Proposed aggregation algorithm

The algorithm 1 shows how the process iterates over multiple window sizes to calculate the slope, aspect, curvature and mean elevation. The input for this method was the DEM data which was stored in the z-array. w_{end} was the largest window size to be evaluated and was used as a stopping condition. Array $xz[i, j]$ consisted of two aggregates. The first aggregate included the $\sum xz$ values showing the shift of z with respect to x-axis. The second aggregate had the $\sum z$ values showing the mean of z obtained from the previous iteration. Both these aggregates consisted of four factors each corresponding to quadrupling across the window sizes. For simplicity, this array could be rewritten as shown in equation 2.13. The same concept was implemented for the $yz[i, j]$ in equation 2.14 followed by $xxz[i, j]$ and $yyz[i, j]$ arrays in equation 2.38 as well by using the appropriate sign convention since the values of $yz[i, j]$ and $yyz[i, j]$ are shifting with respect to y-axis instead of x-axis for $xz[i, j]$ and $xxz[i, j]$. The terms z_{00}, z_{10}, z_{01} and z_{11} represented the z-values for the cells being considered in the bottom-left, bottom-right, top-left and top-right of a quadrangle respectively with the center being at $(0, 0)$. The same sub-script scheme was used for all the variables. The $\frac{w}{4}$ factor was responsible for accounting the coordinate shift in each of these quadrangles.

Data: z, w_{end}; // DEM data and largest window size
Result: $means, slopes, aspects$; // for each w
$xz, yz \longleftarrow zeros$;
$w \longleftarrow 1$;
while $(w < w_{\text{end}})$ **do**
  δ = w
  w *= 2
  foreach $(0 \leq i, j < (size(z) - w + 1))$ **do**
    // $\langle xz \rangle$ as given in Eq. (2.13)
    xz[i,j] = (xz[i][j]+xz[i+δ][j]+xz[i][j+δ]+xz[i+δ][j+δ]
     +δ/4*(-z[i][j]+z[i+δ][j]-z[i][j+δ]+z[i+δ][j+δ])/4
    // $\langle yz \rangle$ as given in Eq. (2.14)
    yz[i,j] = (yz[i][j]+yz[i+δ][j]+yz[i][j+δ]+yz[i+δ][j+δ]
     +δ/4*(-z[i][j]-z[i+δ][j]+z[i][j+δ]+z[i+δ][j+δ])/4
    // $\langle z \rangle$ as given in Eq. (2.39)
    z[i,j] = (z[i][j]+z[i+δ][j]+z[i][j+δ]+z[i+δ][j+δ])/4
    // $\langle xxz \rangle$ and $\langle yyz \rangle$ as given in Eq. (2.38)
    xxz[i,j] = (xxz[i][j]+xxz[i+δ][j]+xxz[i][j+δ]+xxz[i+δ][j+δ]
     +δ/4*(-z[i][j]+z[i+δ][j]-z[i][j+δ]+z[i+δ][j+δ])/4
    yyz[i,j] = (yyz[i][j]+yyz[i+δ][j]+yyz[i][j+δ]+yyz[i+δ][j+δ]
     +δ/4*(-z[i][j]-z[i+δ][j]+z[i][j+δ]+z[i+δ][j+δ])/4
  end
  $means$.add(z)
  $xx = (w*w-1)/12.$
  $xx2 = (w^4 - 5w^2 + 4)/180$
  foreach $(0 \leq i, j < (size(z) - w + 1))$ **do**
    $slopeW[i][j] = \arctan(\text{sqrt}(xz[i][j]^2 + xz[i][j]^2)/xx)$
    $aspectW[i][j] = -\arctan(xz[i][j]/yz[i][j])$
    $curvW[i][j] = xxz[i][j] + yyz[i][j] - 2xx[i][j] * z[i][j]/xx2[i][j]$
    if $(yz[i][j] < 0)$ **then**
      $aspectW[i][j] += \pi$
    end
    else if $(xz[i][j] > 0)$ **then**
      $aspectW[i][j] += 2 * \pi$
    end
  end
  $slopes$.add($slopeW$)
  $aspects$.add($aspectW$)
  $curvatures$.add($curvW$)
end
return $means, slopes, aspects;$

Algorithm 1: Aggregation algorithm.

3. ALGORITHM EVALUATION AND DEM UNCERTAINTY ANALYSIS

3.1. Introduction

Increased availability of high-resolution imagery has exposed limits in the available GIS tools for processing Digital Elevation Models (DEM). Tools that use a 3-by-3 window size for processing these images work well on low resolution DEMs but may produce inconsistent results on high resolution DEM data derived from Lidar point cloud. When topographic features of larger length scales are of interest, high-resolution data have to be resampled to a lower resolution which could produce erroneous predictive models. In the previous chapter a multi-scalar aggregation strategy was proposed that allowed computing topographic features over large windows using the original image size, while having logarithmic run-time efficiency. In this chapter, the algorithm was applied on sample study areas to evaluate its usefulness compared to traditional ArcGIS method. The topographical changes arising due to several length scales were also investigated using semi-variograms from the datasets generated by the proposed aggregation technique and compared to the datasets generated by existing ArcGIS method. Results indicated that the errors encountered during DEM generation decreased at higher window sizes. Results also indicated that the error in modeling was comparatively lower than the values depicted by the ArcGIS based technique.

3.2. Previous work

This chapter builds on the sliding window-based aggregation proposed and implemented in [49]. The evaluation of mean elevation, slope and aspect was done utilizing window sizes $w = 4, 8, 16, 32$ and 64 in [49]. The method described in detail in the previous chapter is summarized in this section.

To evaluate the line of steepest descent used in slope evaluation, a least-squares fit of $z(x,y)$ was performed as shown in equation 2.1. Here x and y represented the horizontal and vertical coordinates of a cell in the DEM respectively. b_0 and b_1 represented the slope along x and y direction respectively. This was followed by minimizing the squared error as in equation 3.1.

$$\langle (z - z_{lin}(x,y))^2 \rangle = \langle (z - b_0 x - b_1 y - c_s)^2 \rangle \tag{3.1}$$

The partial derivatives were calculated with regards to b_0, b_1, and c to minimize the squared error in the evaluation of the steepest descent. Several summation terms such as $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xz$, $\sum yz$ and $\sum z$ required evaluation. Since many terms containing $\sum x$ or $\sum y$ values disappeared due to symmetry of the image as shown in Figure 3.1 the three parameters b_0, b_1, and c were rewritten as in equation 2.8.

	X	Y	XY
Z_1	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$
Z_2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
Z_3	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{4}$
Z_4	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{4}$
	$\sum x = 0$	$\sum y = 0$	$\sum xy = 0$

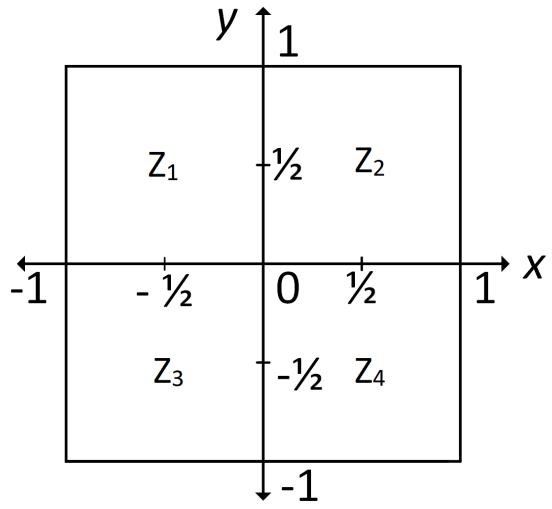


Figure 3.1. Aggregation results of x, y, and xy is zero due to symmetry.

Since the window being considered was a square; $\sum x^2$ was equal to $\sum y^2$ as the coordinates being considered were the same in both the directions. These two terms were evaluated as shown in

equation 2.17 using the concept of power sums as shown in [49]. The slope along two dimensions was obtained using equation 2.10 consisting of $\sum xz$, $\sum yz$ and $\sum x^2$.

To calculate the aspect α, an evaluation in the clockwise direction from the North was performed following the convention used by GIS tools. Using $x = \sin(\alpha)$ and $y = -\cos(\alpha)$, equation 3.2 was derived. The process was discussed in details in the Section 2.3.2. In equation 2.20, the first condition represented the South-West and the South-East quadrant, followed by the second condition that represented the North-West and the third which represented the North-East quadrant of the sub-cell window being evaluated

$$\tan\alpha = -\frac{b_0}{b_1} = -\frac{\sum xz}{\sum yz} \tag{3.2}$$

Curvature was derived by solving a quadratic expression as shown in equation 2.21. This was followed by minimizing the squared error as shown in equation 3.3 and solving partial derivatives with respect to the constants a_{00}, a_{10}, a_{11}, b_{c0}, b_{c1} and c_c . Using these six constants, profile and planform curvatures were derived and used to obtain the mean curvature which was shown in equation 2.33.

$$\langle \big(z - z_{quad}(x,y)\big)^2 \rangle = \langle \big(z - a_{00}x^2 - 2a_{10}xy - a_{11}y^2 - b_{c0}x - b_{c1}y - c_c\big)^2 \rangle \tag{3.3}$$

The sliding window-based aggregation algorithm reuses results from the previous iteration and provides results at multiple scales from a DEM. In Section 3.3 the sliding window-based algorithm was implemented on a DEM. Results of applying slope, aspect and curvature algorithms utilizing the sliding window-based aggregation was compared with the output from ArcGIS using similar window scales. In Section 3.4, the detailed description of DEM errors have been presented where the same study area was used as an example. Finally, in Section 3.5, the ability for these

scales to hold information have been tested using semivariograms at multiple scales for the same study area.

3.3. Materials and methodology

3.3.1. Study area

To evaluate the proposed sliding window based aggregation technique, a DEM of size 1273-by-1273 pixels with a resolution of 5m was used. The study area in Figure 3.2 shows a part of the Bois de Sioux River basin near Tyler in Richland county of North Dakota. The DEM for this location was derived from Lidar provided by International Water Institute [84]. The multispectral image shown in the figure was obtained from Rapid Eye [85].

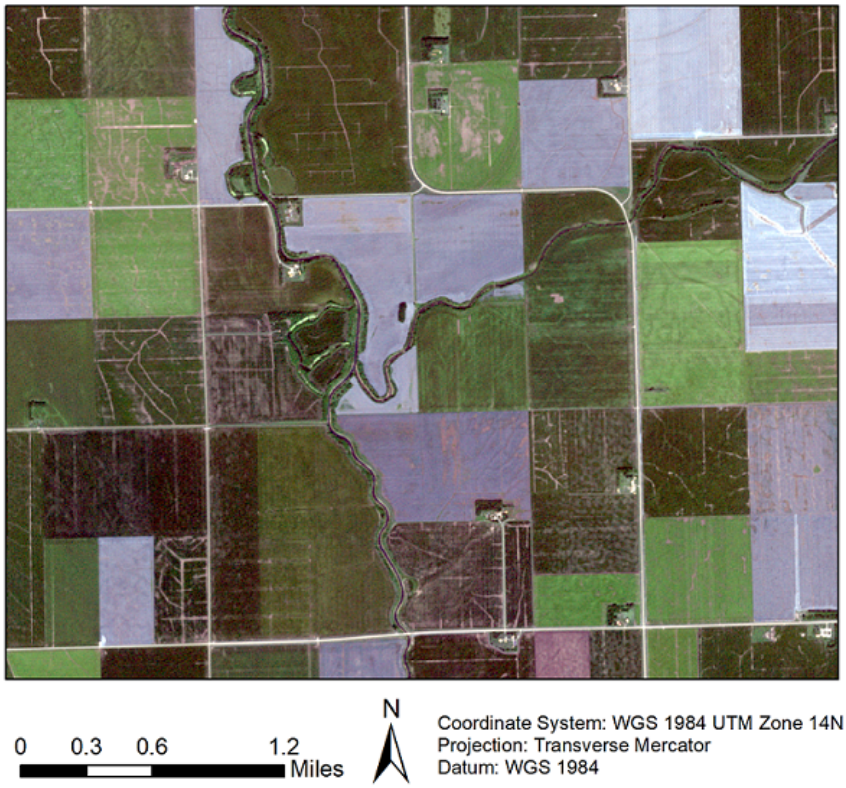


Figure 3.2. Study area.

3.3.2. Sliding window-based aggregation

Figure 3.3 shows the results of the analysis for window-scales 4, 8, 16, 32 and 64 (top to bottom) obtained using the proposed method. The left column showed the sliding window average of elevation values over the respective window-scales. The middle column represented the slope, which was followed by mean curvature on the right. At window-scale $w = 4$, the curvature results showed a lot of noise. For $w = 32$ and 64, noise barely affected the results. The curvature around the river now became prominent and so did the roads around the study area. This pattern was also observed for slope and elevation as the window size increased. The depth of the riverbed which was sharp in lower scales was no longer distinctly visible for $w = 64$. The region surrounding the river basin became more generalized and appeared as a lower elevation region compared to its surrounding. The visualization of these results showed that using any data for $w = 4$ or $w = 64$ for predictive modelling could have a significant impact on the outcome. The choice of the window size to be used for any research was a question that could be justified based on visual interpretation of the spatial data. Hence it is a good practice to view results from multiple length scales to make a logical decision. The proposed algorithm was able to reuse aggregates from previous length scales to derive DEMs of a higher scale, giving the user an opportunity to make the right choice.

3.3.3. Traditional approach using ArcGIS

A similar analysis was done using ArcGIS and is shown in Figure 3.4, using multiple levels of resizing of the data to approximately match the window sizes used in the proposed method. The values for slope and curvature were different compared to the output from the proposed approach in Figure 3.3. This was because the proposed approach applied normalization across the window sizes. Slope and curvature in the first row were evaluated using the original DEM data. In the next row, the DEM was resized, averaging 3-by-3 pixels, followed by implementing the slope and

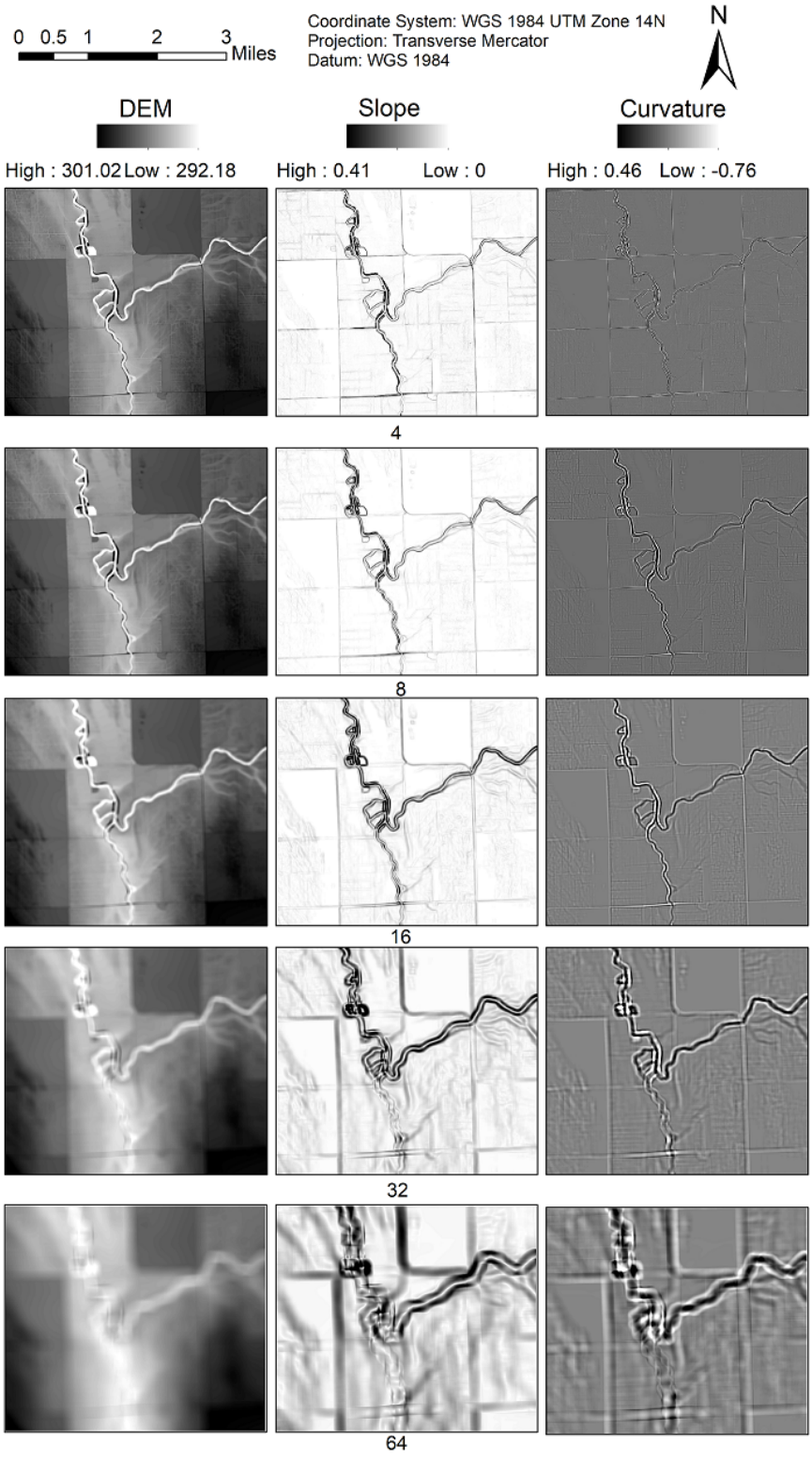


Figure 3.3. Algorithm output for DEM, slope and curvature for proposed sliding window-based aggregation. The proposed approach normalizes outputs for slope and curvature unlike the conventional approach.

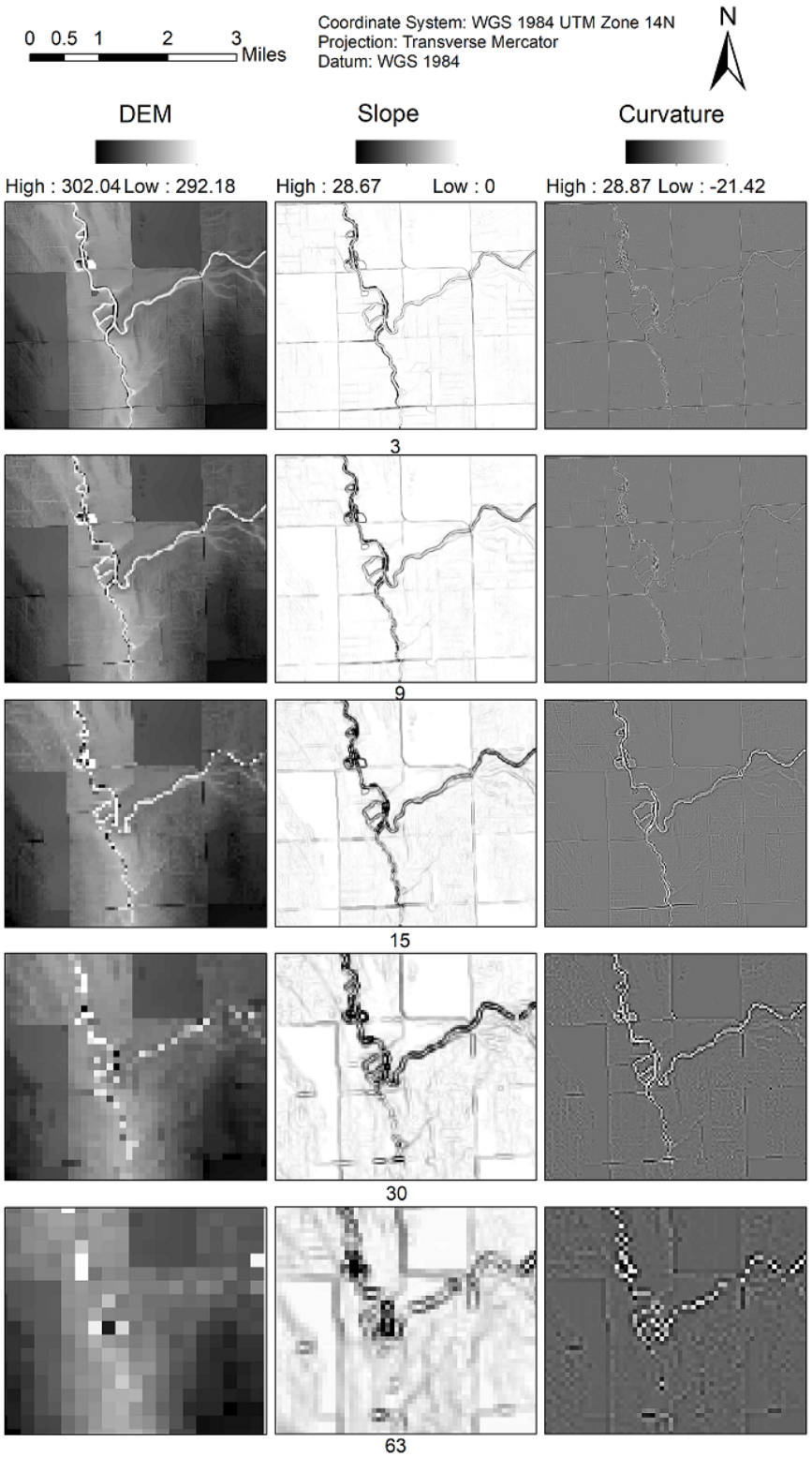


Figure 3.4. Output for DEM, slope and curvature from traditional approach in ArcGIS.

curvature tools available in ArcGIS which also use a fixed 3-by-3 window (nine cells) to derive results. Implementing the tools on a resized DEM of a scale of three, produced results of window-scale $w = 9$, equivalent to window-scale $w = 8$ used by the proposed method. The remaining three rows were obtained by running the slope and curvature tools on DEMs that were resized by a factor of 5, 10 and 21 pixels respectively. This produced slope and curvature results of window-scales 15-by-15, 30-by-30 and 63-by-63 which were equivalent to window sizes 16, 32 and 64 of our proposed method. The pixellation in window-scales $w = 30$ and 63 were very noticeable, and the differences to Figure 3.3 substantial. While the last row in the Figure 3.3 showed a continuous pattern in slope and curvature outputs, the last row in Figure 3.4 showed pixels that almost appear random. The slope information was better than the curvature, but the randomness in pixel values overshadowed the continuous pattern that was present in pixels with a lower window size. Broad regions of slope were visible in the central portion of the image where the river meandered through in $w = 63$. If the study was related to agriculture, the focus on steep inclines next to roads was an issue for slope calculations and it appeared in both the proposed model and the ArcGIS approach across all window sizes. Overall, the sliding window-based aggregation resulted in images that represented features consistently at different length scales.

3.4. Propagation of DEM errors in upscaling

Error analysis was done to derive the extent to which values of the DEM in the current window size showed deviation from the previous scale. Equation 3.4 shows the absolute value of the residual errors incurred to derive the images where y_i was the pixel value at window scale w and $\hat{y}_i$ was the pixel value for scale $\frac{w}{2}$. These errors were visualized in Figure 3.5 for scales $w = 4, 8, 16, 32$, and 64 for proposed method (left) and $w = 3, 9, 15, 30$, and 63 for resampling to lower resolution in ArcGIS. DEM errors accumulated during the resizing process were higher

around the riverbank since there was a lot of change in the elevation values around it. A higher error variation was also observed for the DEMs derived using ArcGIS method at 8.7m than the DEMs obtained from the proposed method at 2.79m.

$$Err_{Abs} = |y_i - \hat{y}_i| \tag{3.4}$$

The DEM data was also used to evaluate three of the most common error metrics, namely mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). Results are shown in Table 3.1. MAE was obtained by taking the sum of the absolute value of all errors and dividing it by the number of pixels present in the image. MSE does this analysis by squaring the difference of both the errors followed by dividing the result with the image resolution. The RMSE was evaluated by finding the square root of the errors derived from mean squared error. Results from both approaches did show a significant difference in the error metrics. While most of these values were less than 1m, error was higher overall for results obtained from ArcGIS method.

Table 3.1. Error analysis across multiple scales derived from the DEM for both approaches.

Proposed Method			
W_{size}	MAE	MSE	RMSE
4	0.036	0.011	0.104
8	0.032	0.008	0.087
16	0.044	0.014	0.121
32	0.051	0.015	0.124
64	0.055	0.01	0.102
ArcGIS Method			
W_{size}	MAE	MSE	RMSE
3	0.0528	0.032	0.179
9	0.1425	0.158	0.398
15	0.11	0.123	0.351
30	0.169	0.207	0.454
63	0.207	0.224	0.473

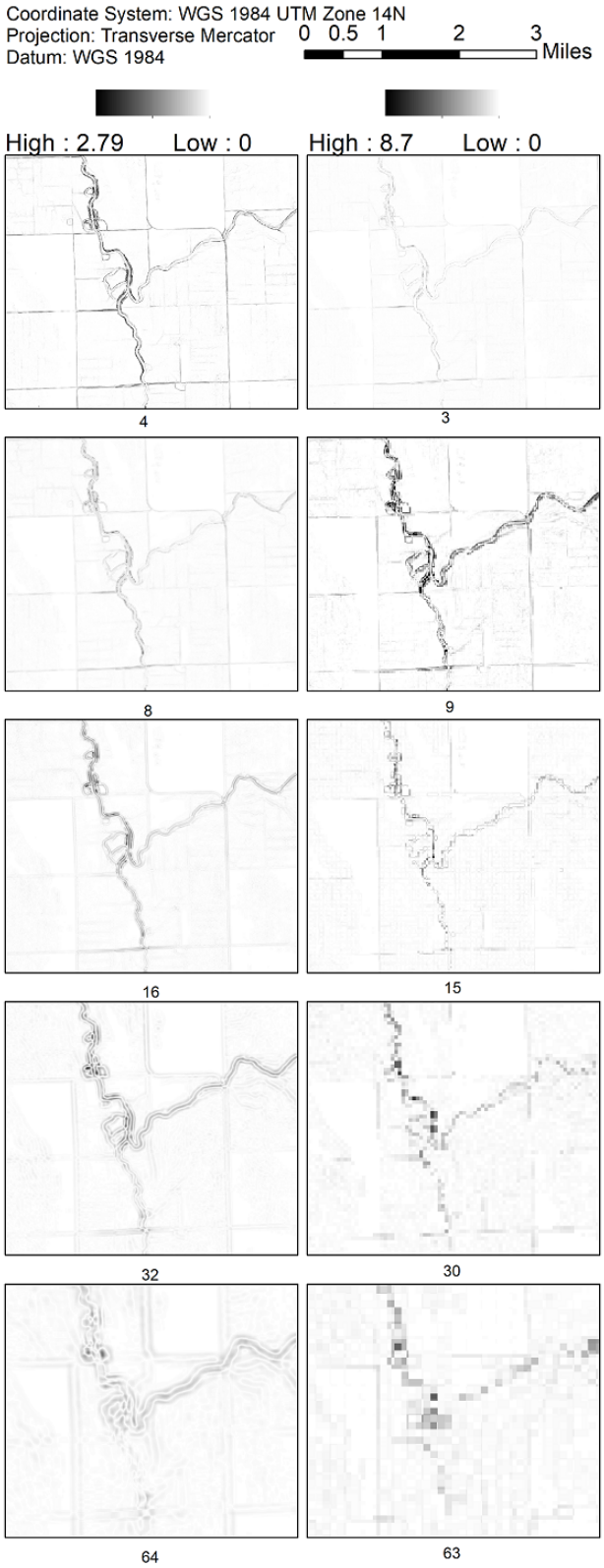


Figure 3.5. Error propagation comparison of sliding window aggregation (left) with ArcGIS method (right) for comparable window scales. A higher value is observed for ArcGIS method of resampling to lower resolution.

The lower error values were justified since the study area was mostly flat with variations near the riverbed. Furthermore, these three error metrics were averaged by the number of pixels in the image, reducing the values to what is observed in the Table 3.1. The RMSE values were larger than the MAE because RMSE is biased towards larger values unlike MAE. Results also showed a trend where the error metrics decreased at a certain window size before increasing again. This behavior was observed at $w = 8$ for the proposed method and $w = 15$ for the ArcGIS method. This trend signified that although the general trend for resizing to a higher length scale was to deviate from actual values and lose information, certain window sizes may be better at explaining spatial auto-correlation making them better suited for analysis. These results provided sufficient motivation to perform a semivariogram analysis which is discussed in the next section.

3.5. Modeling spatial changes in a DEM across multiple window-scales

3.5.1. Semivariogram statistics

Spatial auto-correlation [86] measures the correlation of objects in the DEM separated by a certain distance in space. A semivariogram is an important tool in geostatistics. It is used to model spatial auto-correlation between pixels in the DEM [87]. It is based on the assumption that objects which are nearer are more spatially correlated than the ones farther away [88, 89]. Equation 3.5 [90] is used to model a semivariogram.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left(Z(x_i) - Z(x_i + h)\right)^2 \tag{3.5}$$

Here $Z(x_i)$ and $Z(x_i + h)$ denotes two vectors with elevation (z) values at point locations separated by a lag vector h. The system scans for all pairs of points that are present within a lag vector h and squares their differences. The denominator $N(h)$ denotes that the squares of the differences obtained earlier and is averaged by the number of pixels/points used in the evaluation. The two in the denominator denotes that the expression is a semivariogram and not a variogram.

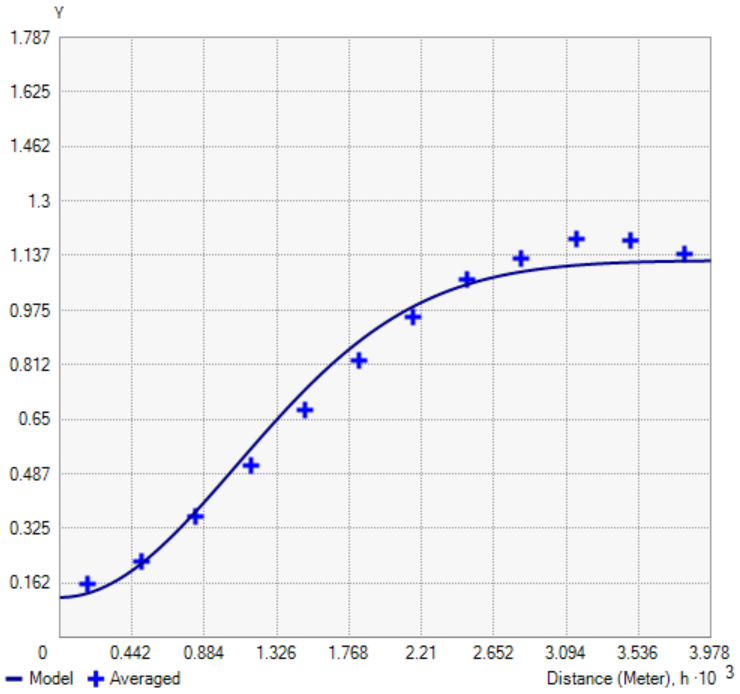


Figure 3.6. An example of a semivariogram plot used to explain the range, nugget and sill.

The process is repeated for multiple lag distances h by increasing its value. Figure 3.6 shows an example of a semivariogram where the x-axis is used to denote the lag distance and y-axis denotes the semivariance. The graph plots the distance with dissimilarity. As the lag distance increases, the semivariance increases since there is less correlation between points that are further apart. Finally, the curve flattens out showing no spatial correlation at larger lag distances. Each plot on the graph is evaluated using multiple points N over a fixed lag distance. The process is repeated using the same points but with a different lag-distance and their results are averaged and shown as the visible plots on the graph. A line of best fit is drawn through the points representing the semivariogram model.

The three measures that are used to define a semivariogram are called the range, nugget and the sill [91]. The range is used to show the lag-distance at which there is no spatial correlation i.e. elevation values at any points being considered are not dependent on another. In the graph,

the range is denoted where the curve starts flattening out. The sill is the maximum variability that is present in the data. This is usually the semivariance observed in the range. Finally the nugget denotes spatial variability in close distances. Usually, points that are adjacent to each other, should have no semi-variance. Hence the model should start from the origin $(0,0)$. However as observed in this graph there is some semivariance even at no lag distance. This semivariance is called the nugget and the phenomenon is termed as the nugget effect. The nugget effect is observed due to several factors like imperfections, errors in the dataset or the sampling technique being used. The total observed variance is obtained by subtracting the nugget from the sill.

3.5.2. Sliding window-based aggregation

To generate spatially autocorrelated random fields, the total of five hundred data points were randomly selected in the study area. Elevation information was extracted for $w = 4, 8, 16, 32$ and 64 corresponding to these points and used for interpolation. Simulated models of the study area were generated using ordinary kriging which relied on information from the semivariogram. To address the required pre-requisites before kriging, data was normalized using the normal score transformation. This transformation used the highest and lowest values in the datasets and matched their ranks to make them normally distributed. Once the dataset was transformed a trend analysis [92] was done to remove any noticeable trends which could introduce a bias [93] during interpolation. Since the riverbank was located in the center of the study area, the trend analysis mostly revealed a U-shaped curve along both the axes. A second order polynomial equation was used to fit the points and reduce the effect of this trend on prediction.

The semivariogram derived for the five window scales was optimized by setting the appropriate lag size to yield best results. Three semivariogram models namely Exponential, Polynomial and Gaussian were evaluated. The Indicative Goodness of Fit (IGF) was used to select the best

model [94]. The IGF denoted the RMSE for fitting the semivariogram as a percentage of sill value. However, IGF also used a weighting factor by virtue of which it gave more significance to results with a smaller lag size than the points which were far apart. Table 3.2 summarizes the IGF across all the models for the semivariograms derived across multiple window scales. It was observed that the exponential model did a better job at ensuring a good fit and reducing the RMSE. The second observation was that the IGF reduced with an increase in the length scale. The highest IGF was observed for $w = 4$ at 0.4199 while lowest was for $w = 64$ at 0.1045. The fact that the IGF reduced with an increase in the length-scale showed that much of the surface variance was being lost by increasing the window size.

Table 3.2. IGF for semivariogram models from proposed sliding window aggregation - lower the better.

	Proposed Method		
W_{size}	Exponential	Polynomial	Gaussian
4	0.4199	0.4239	0.4224
8	0.3669	0.3690	0.3681
16	0.2826	0.2843	0.2836
32	0.1823	0.1848	0.1838
64	0.1045	0.1057	0.1006

The cross-validation results of kriging across the window scales are summarized in Table 3.4. These results were derived by comparing the actual elevation values in the selected points with the predicted values of the semivariogram models. RMSE values indicated that the error rate decreased with an increase in window scale further validating the fact that the DEM does suffer from some degree of generalization due to increasing length scale. All the mean standardized errors were close to zero and it generally reduced with an increase in window size like the RMSE.

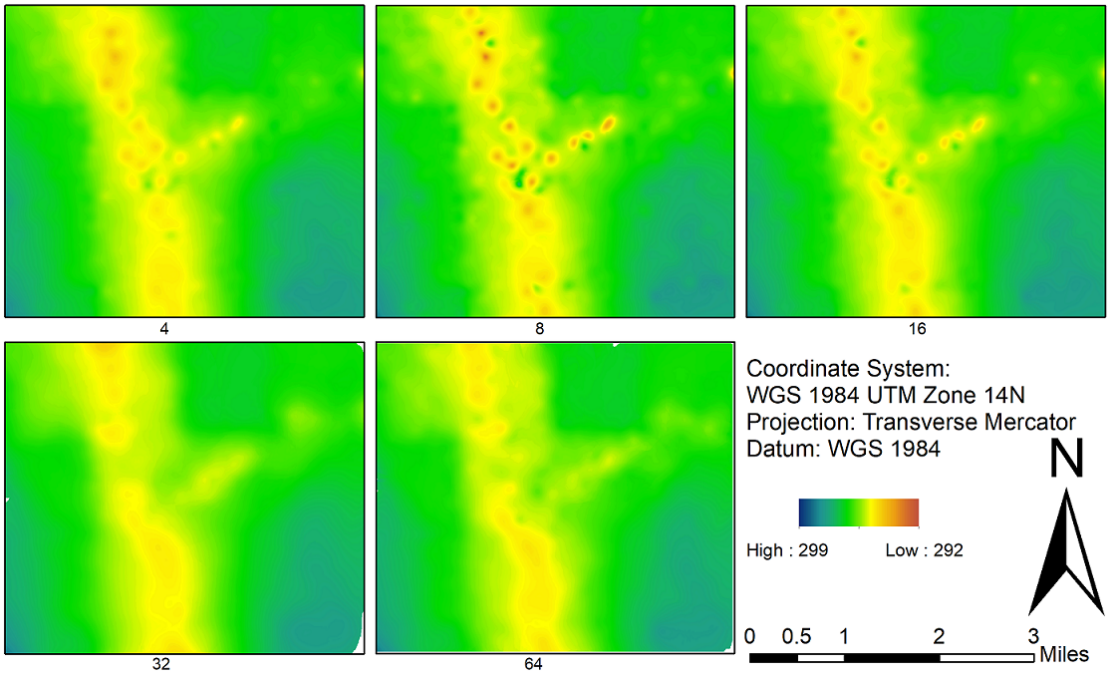


Figure 3.7. Simulated DEMs developed from the semivariogram models using proposed method.

The simulated DEMs for five length scales are shown in Figure 3.7. These results also indicated how the generalization effected the interpolation. DEMs with $w = 4, 8$, and 16 showed patches of low elevation in regions where the river was present. This was mostly caused due to some points falling on the river bed. As the elevation difference between the riverbed and the river bank was comparatively higher and the lower elevation of the riverbed was localized over a narrow region, the interpolation generated lower elevation patches that did not extend over a larger area. As the window scale increased, the generalization effect became predominant in the simulated DEMs for $w = 32$ and 64.

3.5.3. Traditional approach using ArcGIS

The accuracy of generating simulated DEMs was also compared with the results obtained from ArcGIS. Elevation information for the same five hundred points were extracted for comparable window scales, $w = 3, 9, 15, 30$ and 63. The data obtained from ArcGIS method exhibited

similar trends which were removed by fitting a second order polynomial equation. It was also normalized using the normal score transformation before applying ordinary kriging.

Table 3.3. IGF for semivariogram models from traditional approach in ArcGIS - lower the better.

	Traditional method using ArcGIS		
W_{size}	Exponential	Polynomial	Gaussian
3	0.4138	0.416	0.4144
9	0.3791	0.3877	0.3844
15	0.424	0.4264	0.4251
30	0.3786	0.3847	0.3809
63	0.3697	0.3758	0.3741

The IGF values derived from ArcGIS results shown in Table 3.3 were comparable to the one's derived from sliding window aggregation at $w = 3$ and 9. The IGF of $w = 3$ was 0.4138 for the exponential fit which was in fact slightly better than $w = 4$ at 0.4199. As the scale increased, it was observed that the IGF increased dramatically for $w = 15, 30$ and 63 compared to similar results obtained from the proposed model. At $w = 63$ the IGF was 0.3697 for the Exponential fit compared to 0.1045 for $w = 64$. In both approaches, it can be see that the exponential kernel did a better job at defining the simulation model as it produced the lowest IGF values compared to Polynomial and Gaussian kernels. Based on the IGF values, it can be concluded that the semivariogram models derived from ArcGIS did not produce a good fit unlike the proposed method.

The semivariograms were also used to derive simulated models using ordinary kriging. The RMSE derived from cross-validation of five hundred points in Table 3.4 also showed a similar trend as the IGF. The RMSE for $w = 3$ was slightly smaller than $w = 4$. This result correlates to the IGF for being better for $w = 3$. However on increasing the window scales further, the RMSE started increasing and showed higher values compared to the proposed aggregation strategy. The RMSE for $w = 63$ from ArcGIS method was 0.3472 compared to 0.0987 for $w = 64$ from the

proposed aggregation method. Simulated DEMs derived from ArcGIS method are shown in Figure 3.8. Once again the pixellation was very evident in larger window scales such as $w = 30$ and 63. The simulated DEMs also lost pixel density around the river bank on the right side of the image compared to the simulated models derived from window based aggregation.

Table 3.4. Kriging cross-validation comparison for both methods.

Proposed aggregation method		Traditional method using ArcGIS	
W_{size}	RMSE	W_{size}	RMSE
4	0.4214	3	0.414
8	0.3675	9	0.382
16	0.2827	15	0.4241
32	0.1817	30	0.3775
64	0.0987	63	0.3472

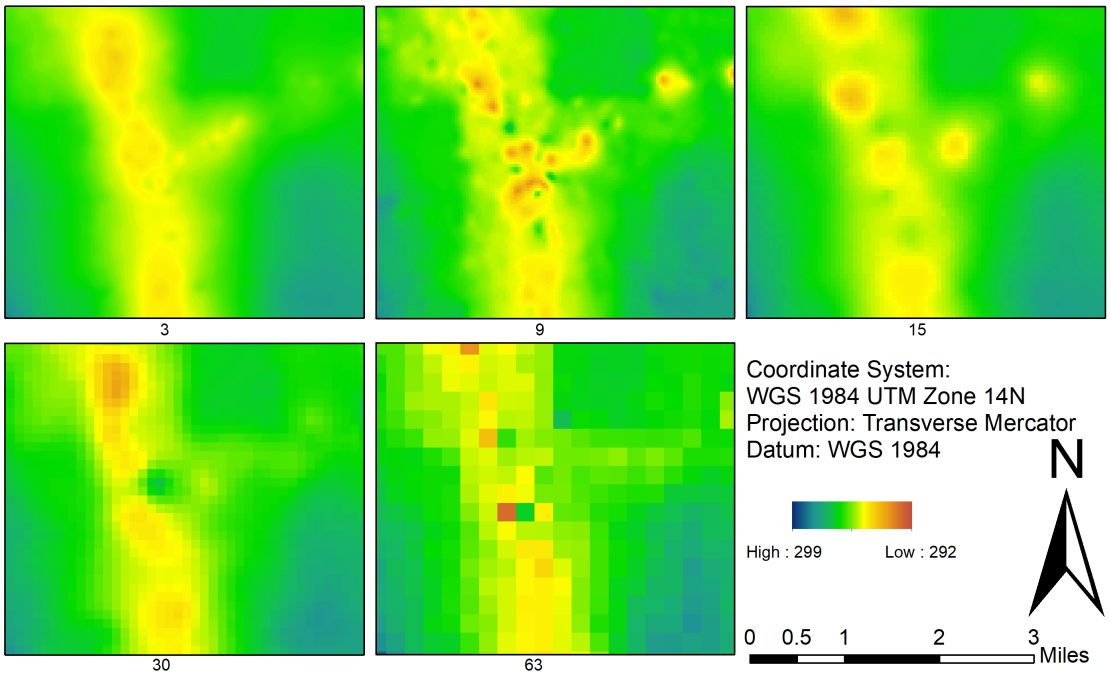


Figure 3.8. Simulated DEMs developed from the semivariogram models using ArcGIS method.

4. COMPARING CLASSIFICATION ACCURACY OF NDVI WITH DEM DERIVED ATTRIBUTES FROM PROPOSED SLIDING WINDOW-BASED AGGREGATION

4.1. Introduction

In the previous chapter, a sliding window based aggregation technique was presented which was able to produce images over multiple scales as well as achieve logarithmic run-time efficiency. Using this aggregation methodology, equations for slope, aspect, and curvature were derived from a DEM. Results obtained from ArcGIS-based method were compared with the output from the proposed method which showed that the ArcGIS based method had a higher RMSE across multiple length scales. The results also suffered from less pixellation compared to ArcGIS output. In this chapter, machine learning models were generated that could be used to predict NDVI using these landform attributes obtained from sliding window-based aggregation.

Classification and clustering are two concepts in data mining that are used for mapping and generating predictive models using GIS and remote sensing software. Clustering or unsupervised learning tries to find similar groups in the data without training data being provided to the system. Classification however requires that a set of training data be provided which contains class labels corresponding to the attribute type. The model is trained using the training data, tested for accuracy and used to classify the testing data. In this chapter, two classification methods were tested to determine their feasibility in generating predictive models, namely Random Forest and Naive Bayes classification.

4.2. Previous work

The second chapter of this dissertation explained the working of a sliding window-based technique and used it to derive an algorithm to calculate the landform attributes slope, aspect and curvature from a DEM. A simple example could be the mean calculation of a DEM. This process was repeated by sliding the window one column each time to the right. It produced a DEM generalized to a 4-by-4 window. Applying another 4-by-4 window on this new result produced a DEM with 16-by-16 window resolution. Each iteration saw points quadrupling even when the window size remained four. Since slope, aspect and curvature also worked on the concept of linear aggregates, the process could also be extended for calculating these topographic variables across multiple window scales. One important factor addressed in this chapter during this window scaling process related to the shift in coordinate system arising from calculating values in a 4-by-4 window. Hence if a window size of w-by-w produced a new window size $2w$-by-$2w$, the corresponding DEM would have a coordinate shift of $\pm(w/4)$ across the entire DEM where w represented the window size. This shift in the coordinate was important to recalculate the slope obtained from aggregating results using least squares as in equation 2.1 and also reducing the mean squared error produced from this shift.

The applicability of sliding window technique for performance efficiency was also argued in [1]. An aggregation algorithm was developed that could generate results across multiple window scales 4-by-4, 8-by-8, 16-by-16, 32-by-32 and 64-by-64. It was used to find the correlation and slope of regression between the Near-Infrared (NIR) and Red bands of a multispectral image of a study area in North Dakota. This approach was able to distinguish shadows cast by trees on the fields by positive correlation results between NIR and Red bands which signified much of the reflectance was being blocked out. As correlation and regression mostly used linear aggregates

such as means and not medians, the results obtained from previous iteration could also be used in the next iteration thereby achieving logarithmic efficiency.

The third chapter discussed in detail the output from the proposed algorithm and compared results with existing resampling to lower resolution technique. Since most of the work in deriving these GIS attributes could be done by fitting a least squares equation to the DEM window, an error propagation model was developed. Results showed that increasing window size caused a loss of topographic detail. These findings supported the research conducted by Wood [75] where he was able to produce windows at a much larger scale and filter out any noise in the dataset that could be produced at a higher resolution.

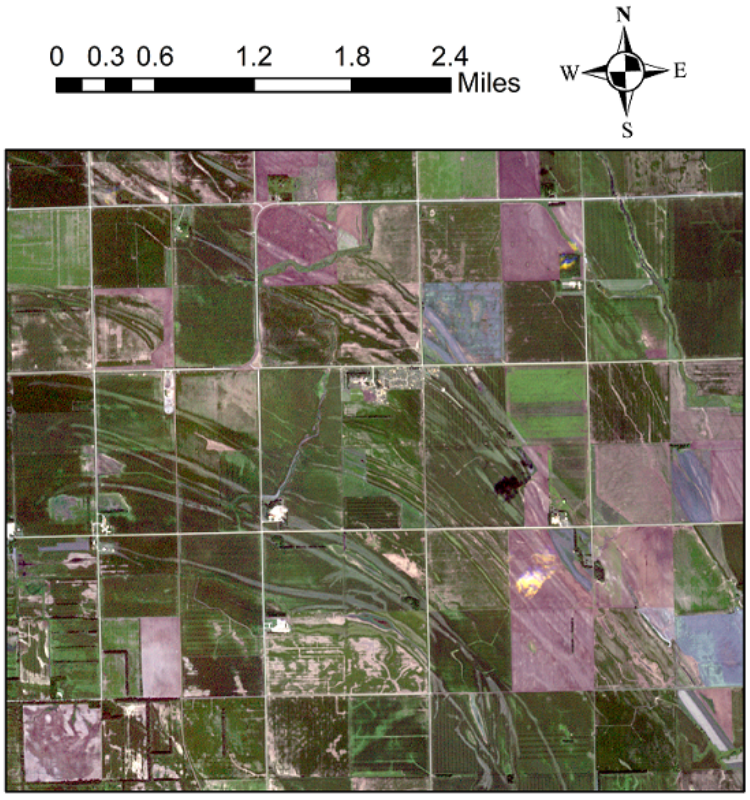


Figure 4.1. Study area.

4.3. Materials and methodology

4.3.1. Study area

Figure 4.1 shows a roughly 3 sq. mile region in Richland county of North Dakota and Roberts County of South Dakota. The DEM was obtained from Lidar 1m resolution data obtained from Red River Basin Decision Information Network [84] while the multispectral image was obtained from Rapid Eye [85] and has a 5m resolution. The Red and NIR bands were used to find the NDVI. This DEM was used to derive the landform attributes slope, aspect, and curvature.

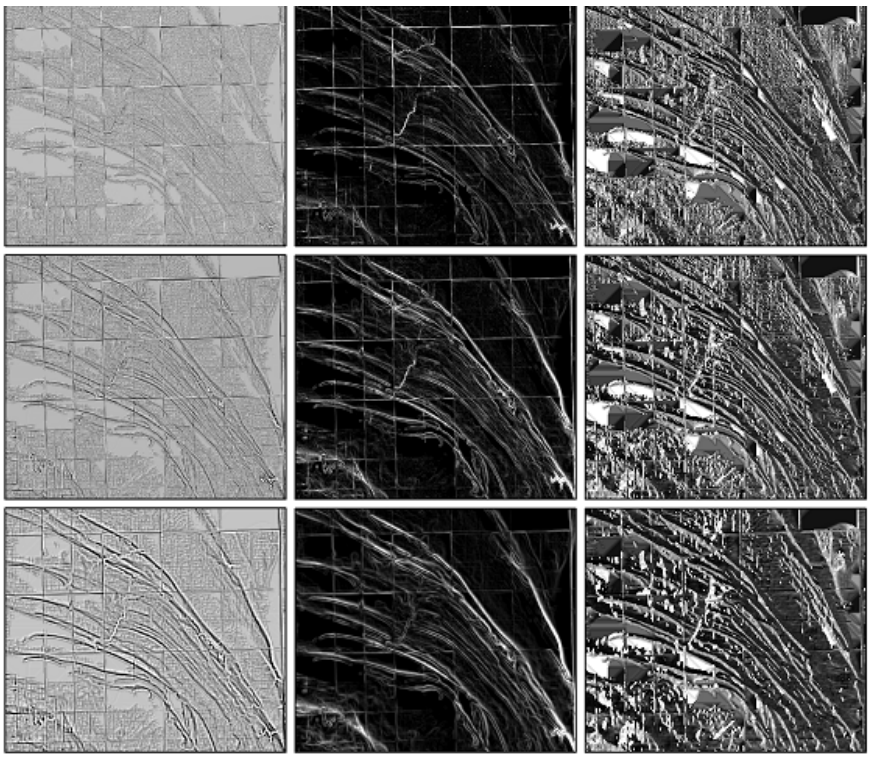


Figure 4.2. Slope, Aspect, and Curvature (left to right) for w = 4, 8, & 16 (top-bottom) from proposed method.

4.3.2. Sliding window-based aggregation

The DEM along with NDVI and the three derived landform attributes slope, aspect, and curvature were obtained using the proposed method for window scales 4-by-4, 8-by-8 and 16-by-

16. The output for slope, aspect and curvature from the proposed methodology is shown in Figure 4.2.

A side by side comparison of NDVI output across both techniques is shown Figure 4.4 with output from proposed method on the left and the ArcGIS results on the right. Results obtained from the window scale $w = 15$ in ArcGIS results showed significant pixellation compared to the results generated by window scale $w = 16$ using the proposed method. Since window scale $w = 15$ was achieved by resampling the DEM by a factor of five, this pixellation was expected considering the fact that each pixel now corresponded to 25m on the ground as opposed to 5m in the beginning. This pixellation observed in ArcGIS output escalated at higher window scales as well and made it difficult to determine a pattern in the field.

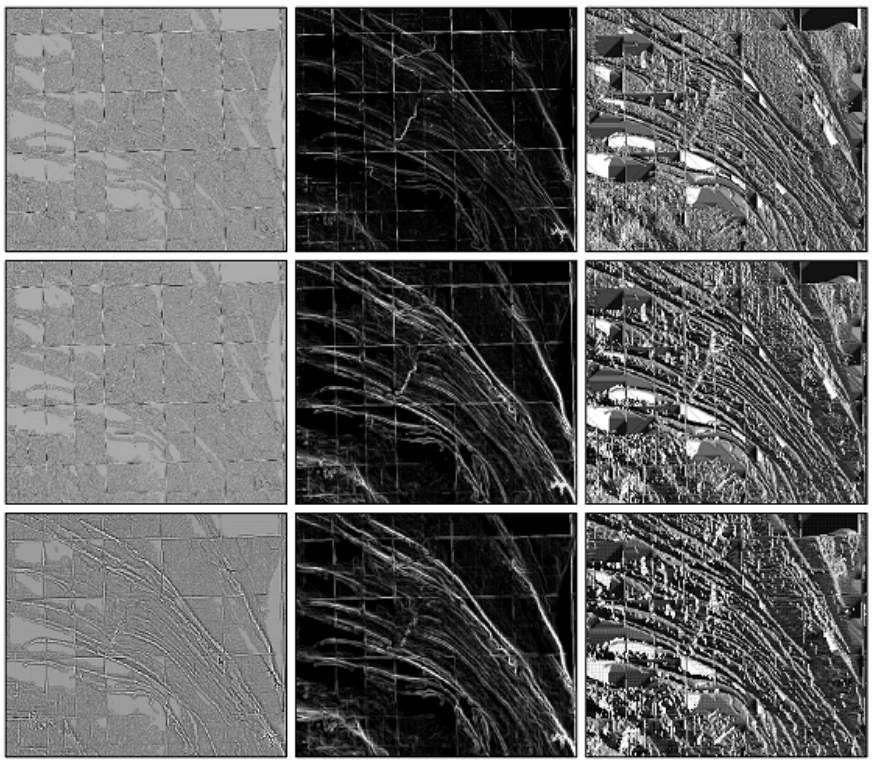


Figure 4.3. Slope, Aspect, and Curvature (left to right) for w = 3, 9, & 15 (top-bottom) from ArcGIS results.

4.3.3. Traditional approach using ArcGIS

Using the traditional method of resampling to a lower resolution, similar results were derived for window sizes 3-by-3, 9-by-9, and 15-by-15 in ArcMap. Since ArcGIS used a fixed 3-by-3 window for evaluating slope, aspect and curvature, the DEM was resampled before deriving these attributes. For a 3-by-3 window the ArcGIS tools were applied on the original DEM. For the 9-by-9 output, the DEM was resampled first by a factor of three and for the 15-by-15 window the DEM was resampled first by a factor of five. The output from the ArcGIS processing is shown in Figure 4.3.

4.4. Results

The objective of this section was to select a data mining model with higher prediction accuracy that could be used for further analysis. Once the DEMs were derived, results were exported to R for analysis. It was filtered based on positive curvature values to reduce the sample size as the original DEM contained 3,126,196 pixels with a resolution of 1949-by-1604. Outliers in the DEM and NDVI could be seen as unusually high or low values with respect to their surroundings. To mitigate their impact on the entire dataset, results were aggregated on NDVI assuming that the majority of values would overshadow the effect of these outliers. The dataset was partitioned using 60:40 rule where 60% was used for training and 40% for testing. The NDVI dataset was converted from continuous to categorical. Six classes were derived using equal interval classification. Table 4.1 shows the range of NDVI values. The range started from 0.3 as any value below it was mostly barren or uncultivated fields. This was followed by the application of two classification algorithms, Naive Bayes [95] and Random Forest on the training dataset.

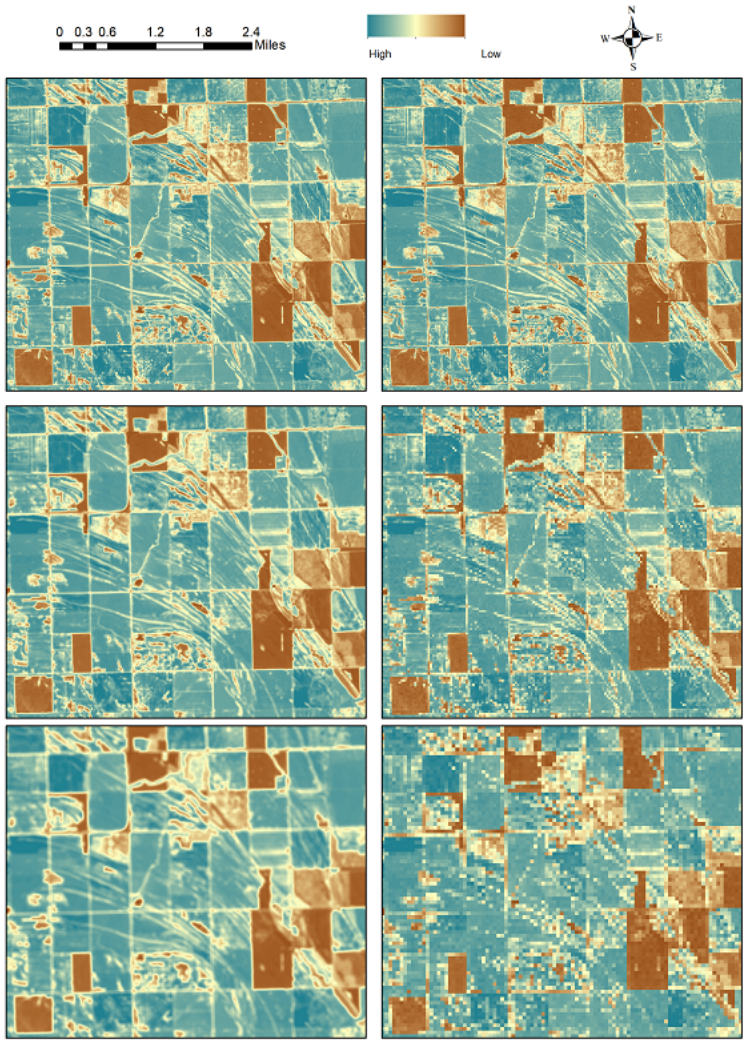


Figure 4.4. NDVI - proposed method (left) & ArcGIS method(right) for comparable window sizes.

4.4.1. Naive Bayes based classification

This classification technique is based on Baye's theorem and conditional probability. The training data contained records corresponding to each class that had been selected previously. The mean and covariance of the records in each class was calculated. In training data, the value of each record along with the mean and covariance was used in the Gaussian distribution to estimate the probability. Each record contained as many probability values as the number of classes. The record was assigned to the class with the highest probability. The process was similar for records

Table 4.1. NDVI classes used to generate predictive models using Naive Bayes and Random Forest.

NDVI Values	Class
0.3 - 0.4	1
0.4 - 0.5	2
0.5 - 0.6	3
0.6 - 0.7	4
0.7 - 0.8	5
0.8 - 0.9	6

with multiple attributes, where instead of having one attribute per record, there were four attributes slope, aspect, elevation, and curvature. The mean and covariance of all the records were calculated in the training data set and used for evaluation in the testing data set. In equation 4.1 [96], the mean and standard deviation were represented using μ_m^k and σ_m^k respectively. x_m represented the attribute for the record value and c_k represented the cluster in which the probability density was being evaluated. The c_k increased with the number of classes being considered [96].

$$P(x_m|\ c_k) = \frac{1}{\sqrt{2\pi\sigma_m^{k2}}} exp^{-\left(x_m-\mu_m^k\right)^2/2\sigma_m^{k2}} \tag{4.1}$$

$$P(\mathbf{x}|\ c_k) = P(x_1, x_2, x_3,\ \ldots\ ,\ x_d \mid c_k) = \prod_{m=1}^{d} P(x_m|\ c_k) \tag{4.2}$$

Table 4.2. Accuracy for Naive Bayes classification on training dataset.

Proposed Method		
Window Size	OOB accuracy	Kappa
4	73.98%	0.6745
8	71.49%	0.6427
16	68.31%	0.6027
ArcGIS method		
Window Size	OOB accuracy	Kappa
4	49.65%	0.3587
8	44.7 %	0.2842
16	44.75 %	0.2018

Table 4.3. Accuracy and confusion matrices for Naive Bayes classification on testing dataset using proposed method.

Proposed Method Confusion Matrix								
W=4	Reference NDVI						Total	
Pred	1	2	3	4	5	6		
1	261	73	0	0	0	0	334	Accuracy
2	61	225	36	0	0	0	322	73.57%
3	2	77	339	57	1	0	476	
4	0	1	29	228	54	0	312	Kappa
5	0	0	0	107	333	0	440	0.67
6	0	0	0	0	0	0	0	
Total	324	376	404	392	388	0	1884	
W=8	1	2	3	4	5	6	Total	
1	189	73	5	0	0	0	267	Accuracy
2	115	244	45	0	0	0	404	70.78%
3	10	56	325	48	0	0	439	
4	0	0	29	227	43	0	299	Kappa
5	0	0	0	119	330	0	449	0.634
Total	314	373	404	394	373	0	1858	
W=16	1	2	3	4	5	6	Total	
1	183	41	12	19	0	0	255	Accuracy
2	34	235	84	0	0	0	353	67.21%
3	77	93	296	59	0	0	525	
4	11	2	12	193	36	0	254	Kappa
5	0	1	0	124	333	0	458	0.589
Total	305	372	404	395	369	0	1845	

In this study the classes were treated as the NDVI categorical values deduced earlier. To estimate the likelihood of a point $\mathbf{x}$ belonging to any class c_i the joint probability of all dimensions belonging to $\mathbf{x}$ was calculated as in equation 4.2 [96]. The model selected the class that maximized posterior probability for a particular $\mathbf{x}_i$. Each $\mathbf{x}_i$ contained four attributes slope, elevation, aspect and curvature which were used to calculate the posterior probability. Thus, the dimension of each $\mathbf{x}$ was four and the number of classes k was 5 or 6 depending on the results obtained either from proposed model or ArcGIS method. Table 4.2 shows the accuracy and kappa values for results on the training dataset. Results from window-based aggregation for $w = 4, 8$ and 16 showed a higher accuracy rate compared to the classification on the ArcGIS results. For results obtained

from the proposed method shown in Table 4.3, the accuracy on testing data ranged from 73.57%, 70.78% and 67.21% for window scales 4, 8 and 16 respectively. It was observed that the accuracy decreased with an increase in the window scales.

Naive Bayes classification was also applied on results derived from ArcGIS for $w = 4, 8$ and 16. The accuracy of the model was comparatively lower when data from ArcGIS was used. This was mostly due to the pixellation observed in the images where vital information was lost due reduction in resolution. Table 4.4 shows that the accuracy on testing data ranges from 47.81%, 42.15% and 40.58% for window scales 3, 9 and 15 respectively.

Table 4.4. Accuracy and confusion matrices for Naive Bayes classification on testing dataset for ArcGIS results.

ArcGIS results Confusion Matrix								
W=3	Reference NDVI						Total	
Pred	1	2	3	4	5	6		
1	38	12	2	1	8	1	62	Accuracy
2	74	144	70	4	30	0	322	47.81%
3	47	133	178	67	28	0	453	
4	31	50	108	290	100	0	579	Kappa
5	47	47	26	43	211	0	374	0.337
6	0	0	0	0	12	1	13	
Total	237	386	384	405	389	2	1803	
W=9	1	2	3	4	5	6	Total	
1	7	16	7	1	5	0	36	Accuracy
2	42	100	67	9	28	0	246	42.15%
3	52	102	119	38	70	0	381	
4	22	65	113	342	102	0	644	Kappa
5	37	71	50	20	100	0	278	0.249
Total	160	354	356	410	305	0	1585	
W=15	1	2	3	4	5	6	Total	
1	1	0	2	0	0	0	3	Accuracy
2	2	7	8	1	2	0	20	40.58%
3	13	57	58	35	16	0	179	
4	33	109	119	291	115	0	667	Kappa
5	7	22	36	16	48	0	129	0.154
Total	56	195	223	343	181	0	998	

4.4.2. Random Forest based classification

Random Forest classifier [97, 98] was chosen for this study as it is an ensemble learning technique which combined the results obtained from multiple decision trees to classify a record. This offered an advantage over a single decision tree classifier like the ctree [99] which mostly aggregated results based on one decision tree. Unlike most decision trees which performed a univariate split, the Random Forest model used its $mtry$ value that combined multiple attributes to perform an efficient split. This process increased the chance of exploring results with fewer nodes as well.

Table 4.5. Accuracy and confusion matrices for Random Forest classification on testing dataset using proposed method.

Proposed Method Confusion Matrix								
W=4	Reference NDVI						Total	
Pred	1	2	3	4	5	6		
1	233	58	1	0	1	0	293	Accuracy
2	90	254	56	0	0	0	400	74.47%
3	1	63	296	55	1	0	416	
4	0	0	50	286	52	0	388	Kappa
5	0	1	1	51	334	0	387	0.68
6	0	0	0	0	0	0	0	
Total	324	376	404	392	388	0	1884	
W=8	1	2	3	4	5	6	Total	
1	194	98	7	0	0	0	299	Accuracy
2	108	224	52	0	0	0	384	72.07 %
3	11	51	313	52	0	0	427	
4	1	0	32	278	43	0	354	Kappa
5	0	0	0	64	330	0	394	0.65
Total	314	373	404	394	373	0	1858	
W=16	1	2	3	4	5	6	Total	
1	187	33	27	18	1	0	266	Accuracy
2	51	260	97	0	0	0	408	71.17%
3	55	76	249	39	0	0	419	
4	12	3	31	290	41	0	377	Kappa
5	0	0	0	48	327	0	375	0.638
Total	305	372	404	395	369	0	1845	

Out-of-bag (OOB) accuracy is a method to ascertain the accuracy of the training model before performing cross-validation with the test dataset. The OOB accuracy corresponding to window scales 4, 8 and 16 were 75.14%, 75.72% and 71.86% respectively as shown in Table 4.7. The accuracy of Random Forest on the testing data was also high as shown in Table 4.5. It varied from 74.47% in window scale 4 to 71.17% in window scale 16. This proved that the results generated by proposed method were better at drawing predictive models that could be used effectively to quantify relationships between multi-variate attributes in the geo-spatial domain.

Table 4.6. Accuracy and confusion matrices for Random Forest classification on testing dataset from ArcGIS results.

ArcGIS results Confusion Matrix								
W=3	Reference NDVI						Total	
Pred	1	2	3	4	5	6		
1	37	25	19	3	16	0	100	Accuracy
2	83	162	101	20	40	0	406	46.31%
3	43	120	160	74	41	0	438	
4	20	27	77	232	48	0	404	Kappa
5	54	52	27	76	244	2	455	0.319
6	0	0	0	0	0	0	0	
Total	237	386	384	405	389	2	1803	
W=9	1	2	3	4	5	6	Total	
1	17	34	11	4	11	0	77	Accuracy
2	52	118	87	18	54	0	329	41.77%
3	42	110	126	57	65	0	400	
4	16	44	81	286	60	0	487	Kappa
5	33	48	51	45	115	0	292	0.251
Total	160	354	356	410	305	0	1585	
W=15	1	2	3	4	5	6	Total	
1	6	7	10	0	2	0	25	Accuracy
2	9	25	28	14	10	0	86	38.38 %
3	16	67	61	56	34	0	234	
4	23	79	91	242	86	0	521	Kappa
5	2	17	33	31	49	0	132	0.151
Total	56	195	223	343	181	0	998	

A Random Forest model was also applied on the results generated by ArcGIS. The derived model also contained a handful of pixels where NDVI attribute ranged from 0.8-0.9. These pixels

Table 4.7. Out of bag (OOB) accuracy and GINI index on training dataset for Random Forest classification.

# trees	Proposed method					
	Window size	OOB accuracy	GINI impurity decrease			
			Curv	Slope	Aspect	Elev
500	4	75.14	831.78	594.25	356.96	492.07
	8	75.72	939.86	477.62	341.87	483.05
	16	71.86	735.41	465.40	448.33	570.13
	ArcGIS method					
	Window Size	OOB accuracy	GINI impurity decrease			
			Curv	Slope	Aspect	Elev
	3	46.28	514.72	538.08	489.52	618.86
	9	38.47	406.64	458.75	475.19	525.45
	15	39.77	261.33	268.99	290.46	319.35

were not present at higher window scales. These outliers were not visible in the proposed model which had five classes instead of six to begin with. It was likely that the averaging effect of sliding window removed such outliers. Due to resampling to a lower resolution in ArcMap, the number of observations reduced with increasing window size, while it remained fairly consistent in the proposed method. OOB accuracy estimates from training data in Table 4.7 were 46.28%, 38.47% and 39.77% respectively for window scales 3, 9 and 15 for results from ArcGIS. The overall accuracy along with the generated confusion matrix using the Random Forest model when evaluated on the test dataset is shown in Table 4.6. The model generated lower accuracy than the proposed aggregation which also decreased further as the window scale increased.

The Gini index [100] was further analyzed corresponding to training data and the results are shown in Table 4.7. Gini index showed the purity of the split done on the tree. A pure split ensured reduced number of nodes and computation. A higher Gini index showed that the attribute played a significant role in the model prediction. The model generated by ArcGIS showed consistent and low Gini values across all the attributes. However the Random Forest model from proposed method showed very high Gini values for the attribute curvature compared to the other three. This showed

that curvature played a crucial role in the analysis. Curvature was mostly followed by elevation and slope while aspect has the least effect on the accuracy of the model.

4.5. Discussion

4.5.1. Naive Bayes based classification

Based on the classification results it can be sufficiently concluded that the results obtained using the proposed method were more accurate in generating a Naive Bayes classification model. The accuracy for $w = 4$ on testing data was 73.57% compared to 47.81% for $w = 3$. It was also observed that the classification model derived from both methods showed a decrease in accuracy as the window size increased. For $w = 16$ the accuracy of Naive Bayes on testing data is 67.21% whereas for $w = 15$ it was 40.58%. With an increase in the window scale there was a generalization effect as discussed in Chapter 3 which reduced the accuracy of the output to a certain extent. Another interesting observation was that the accuracy derived using Naive Bayes was less than the Random Forest model. The prediction accuracy for $w = 4$ from Random Forest model on testing data was 74.47% and 71.17% for $w = 64$. One factor that may account for lower accuracy of the Naive Bayes model could be its underlying assumption. Since Naive Bayes was derived on conditional probability which assumed that the attributes slope, elevation, aspect and curvature were independent of each other. Random Forest however can use a combination of these attributes by changing its $mtry$ value which provided the flexibility of discovering any dependencies between the attributes that may be better in explaining the model. Another factor could be the nature of the Random Forest model. Since Random Forest built multiple decision trees by changing its $mtry$ values and then polled their results to come to a conclusion, it offered an advantage over Naive Bayes which was only run once to create the model.

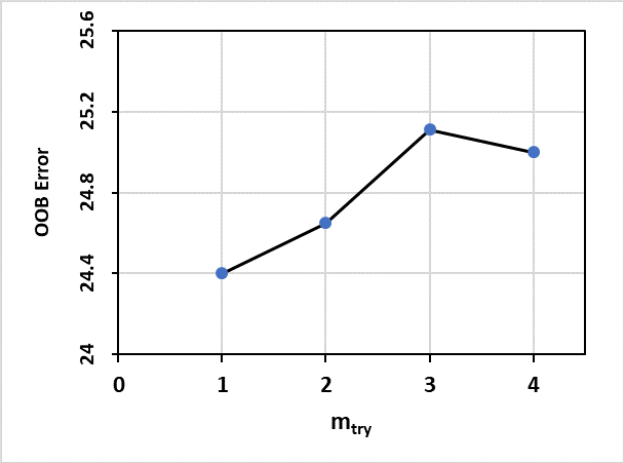


Figure 4.5. Out of bag (OOB) error variation on training dataset from Random Forest classification with mtry for $w = 4$.

4.5.2. Random Forest based classification

By default the Random Forest model used *mtry* value which was square root of the number of predictor variables. Since there were four predictor variables, the default *mtry* was two. Tuning the Random Forest model was done to see if a higher accuracy could be achieved on the results obtained for $w = 4$. The comparison of OOB error with the *mtry* value is shown in Figure 4.5 . The OOB error increased with an increase in mtry for $w = 2$ and 3 and then reduced for $w = 4$. Results indicated that for this classification, a univariate split performed better than a multivariate split.

A plot for the relative error of each class against the number of trees for training data is shown in Figure 4.6. Five lines with varying colors and a black line in the middle can be observed. The black line resembled the overall OOB error close to 25% which matched our recorded OOB accuracy of 75.14% in the training dataset for $w = 4$ in Table 4.7. NDVI class interpretation errors included 26.07% for class 1(red), 32.85% for class 2 (green), 28.02% for class 3 (blue), 22.53% for class 4 (aqua) and 14.10% for class 5 (purple). It was observed that the model tapered to a constant value after three hundred trees. Hence, using more than three hundred trees would be a wastage of memory resources. The model was re-run on training dataset with mtry= 1 signifying only

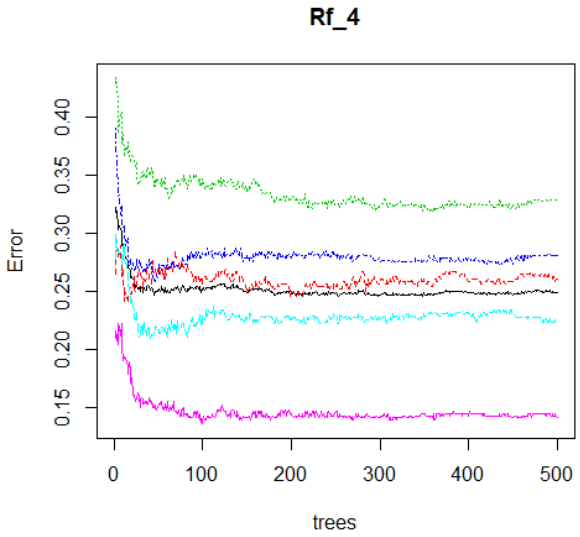


Figure 4.6. Out of bag (OOB) error variation of training dataset with NDVI classes for $w = 4$. Each line corresponds to an NDVI class and the corresponding error encountered during prediction in that class

one predictor variable used for the splits and a reduced tree count. The OOB accuracy increased slightly from 75.14% to 75.21%. On using the new training model on the test dataset, the accuracy of testing data for $w = 4$ increased from 74.47% to 75.16%.

5. ACCURACY ESTIMATION USING RANDOM FOREST BASED REGRESSION MODEL AND DATA VISUALIZATION USING PARTIAL DEPENDENCE PLOTS

5.1. Introduction

In the previous chapter, two classification models namely Random Forest and Naive Bayes were compared to establish which was more accurate for predictive modelling. Slope, aspect, curvature and elevation information were derived from proposed aggregation and traditional method. The results were used to predict NDVI. The classification results from Naive Bayes and Random Forest showed that the data derived from proposed aggregation was more accurate in predicting NDVI than the data derived from ArcGIS-based approach. It was also established that Random Forest was better at developing predicting results compared to Naive Bayes.

In this chapter Random Forest was applied to build a regression based predictive model to compare results from sliding window based aggregation with the ArcGIS output. Regression model offered the advantage of using the entire dataset without segregating results into any classes. This reduces any error arising from the grouping schemes being used for the classes. Random Forest was chosen for this study as it is an ensemble learning technique which combined the results obtained from multiple decision trees and performed better in our previous classification tests. Relationship between the land form attributes were also visualized using partial dependence plots. Results derived in several other study areas having depressions were also investigated.

5.2. Previous work

Several techniques are available that could be used to visualize model performance. Some common ones reported in [101] are partial dependence plots [102], multi-dimensional scaling [103]

and conditional density estimates [104]. Partial dependence plots have been used as an effective tool to visualize the output derived from several machine learning techniques such as Random Forest [105] and XGBoost [106]. In [105], classification accuracy of Random Forest was compared to three other classifiers to identify invasive plant species in Lava Beds National Monument in California, presence of a particular lichen species in Pacific North-West and bird nesting sites Uinta Mountains in Utah. The authors in [105] also compared the cross-validation accuracies of Random Forest with classification trees, logistic regression and linear discriminant analysis and established that the Random Forest approach achieved the highest accuracy among all the statistical classifiers for the three study areas. Partial dependence plots were also used in [105] to show the relation between bird nesting species with two predictor variables differing in the diameter of nesting trees and to study the relation of precipitation, elevation, and age of conifers, to accurately identify the presence of three lichen species in the Pacific North-West. In [106], authors used fifty two features derived from satellite and land-use data to build machine learning models that can improve the accuracy of satellite derived aerosol optical depth (AOD) products by reducing the error for effective air pollution modelling. Relative azimuth angle was a predictor variable with highest importance in the training dataset obtained from Aqua satellites and was used to derive partial dependence plots with respect to the AOD in [106].

A partial dependence plot allows visualization of a machine learning model by showing the marginal effect that one feature has on the outcome from a set of two or three features [107]. Usually a partial dependence plot is created by using a set of two features that have a significant impact on the predicted outcome. Equation 5.1 [102] shows the partial dependence function for a machine learning model being implemented.

$$\hat{F}(\mathbf{X}) = \hat{F}(y_a, y_b) \tag{5.1}$$

Here y_a and y_b denotes subsets from a complete set of records belonging to a feature vector $\mathbf{X}$. In other words, y_a and y_b are complements of each other. The $\hat{F_{\mathbf{X}}}$ depends on both these subsets. In some scenarios, a partial dependence from one subset of features can be used to condition the other subset. Equation 5.2 [102] is a function that is dependent on records in y_a and can be used to explain y_b.

$$\hat{F}_{y_a}(y_a) = \hat{F}(y_a \,|\, y_b) \tag{5.2}$$

There are several instances where, it is relatively difficult to plot a trend of an outcome with respect to all the records of a corresponding feature [102]. In such a scenario, a subset from the set of records could be used to explain the relative dependence of the outcome to the corresponding attribute by using an averaging function. This relation is depicted in equation 5.3 [102].

$$\bar{F}_a(y_a) = E_{y_b}[\hat{F}(\mathbf{x})] = \int \hat{F}(y_a, y_b)\mathbf{P_b}\mathbf{dy_b} \tag{5.3}$$

Here $\mathbf{P_b}\mathbf{dy_b}$ is the marginal probability density of y_b. Like conditional probability it is assumed that the feature y_a is independent of y_b. This approach can be used as an indication of the relation the outcome shares with the corresponding features.

In this chapter, the predictive models were derived from the results of sliding window-based aggregation using Random Forest based regression. Results were also visualized using partial dependence plots to show the relation between the attributes and the outcome. Partial dependence plots were derived in R [44]. Both 2-D and 3-D plots were studied to explain the inter-dependence of the most significant attributes.

5.3. Materials and methodology

5.3.1. Study area

For this analysis, a study area was chosen which spanned a region of Richland county in North Dakota and Roberts county in South Dakota as shown in Figure 5.1 was used. The DEM

was derived from Lidar point cloud data and provided by International Water Institute, Fargo [84]. The multispectral images were obtained from RapidEye [108, 85] and had images captured in five spectral bands of which Red and Near-Infrared were used. The grid size was 1949-by-1604 pixels of 5m spatial resolution. The Red and Near-Infrared (NIR) bands were preprocessed along with the DEM before analysis [109]. NDVI was derived from the Red and the NIR bands using equation 5.4 [110].

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{5.4}$$

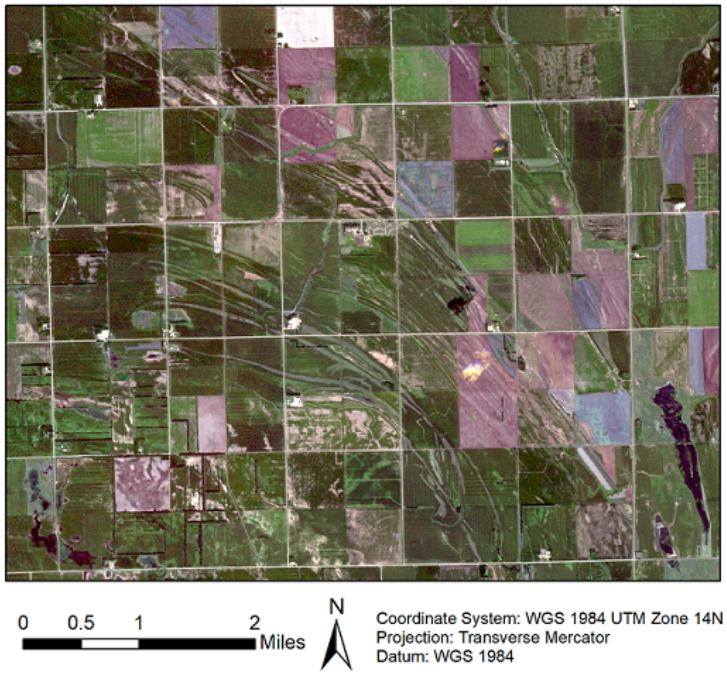


Figure 5.1. Study area.

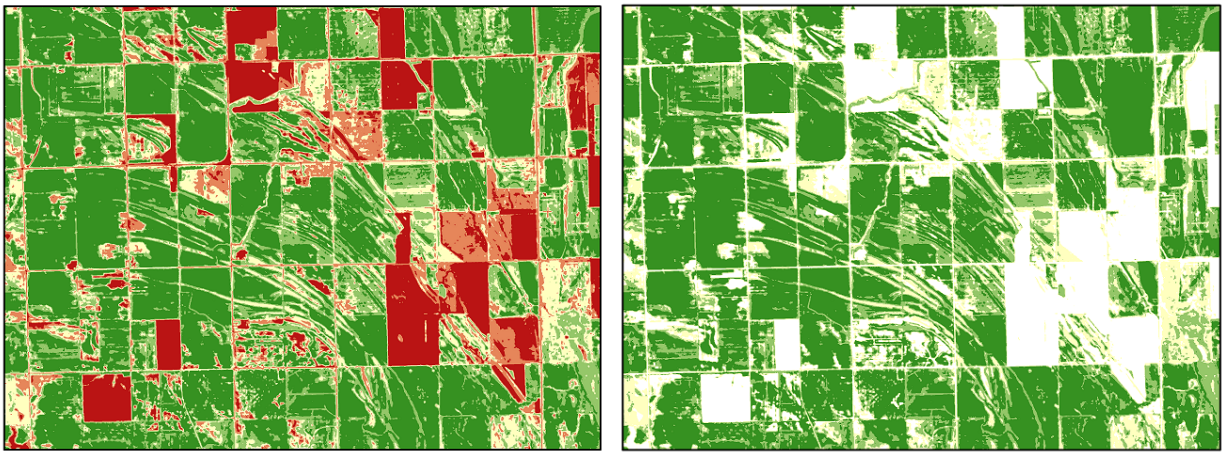
NDVI is a strong indicator of vegetation health. Its value ranges from -1 to 1 [111]. A higher NDVI value denotes good while low NDVI denotes poor vegetation health [111]. This index was obtained using the Red and NIR band of a multispectral image. Healthy vegetation

absorbs a lot of light in the Red spectrum and reflects the light in the NIR spectrum [111]. Using equation 5.4, it can be observed that the NDVI is higher when the reflectance of the NIR is higher. Negative NDVI usually correlates to water [112] and urban settlements have NDVI closer to zero [113]. There were not enough observations with negative NDVI values in our dataset as it mostly contained agricultural fields. There existed some heterogeneity in the datasets based on difference in crop planting and harvesting period arising from farmers shifting their harvesting to an earlier date or keeping the land barren resulting in low NDVI values. Since these regions by no means indicated soil health, they were excluded from the calculation by using cut-off values for NDVI across several window sizes. The cut-off was implemented using the Jenk′s Natural Breaks algorithm [72]. Since this algorithm divided the dataset based on significant differences (breaks) in data values, it was the most appropriate algorithm for the task. Figure 5.2a shows the NDVI of the study area before application of the algorithm. Green represented a higher NDVI while areas in Red represented fields left barren or fallow. For a window of size four, the algorithm selected 0.32 as the cut-off NDVI to distinguish barren lands from the cultivated ones. The resulting NDVI is shown in Figure 5.2b. Using the NDVI raster as a mask; slope, elevation, aspect and general curvature values were extracted using the ′Extract by Mask′ toolset followed by the ′Extraction of Values to Points′ [114] to convert the raster data to a tabular format. This data was imported to R [115] for further analysis. The process was repeated across several window sizes.

5.3.2. Sliding window-based aggregation

Using the proposed sliding window-based aggregation technique [1, 49]; slope, aspect, curvature, and elevation were obtained for window scales 4-by-4, 8-by-8, 16-by-16, 32-by-32 and 64-by-64. Figure 5.3a shows the output for curvature, slope and aspect (left to right) as the window sizes doubled in each iteration (top to bottom). While comparing both methods, the effect

of resampling to a lower resolution was visible on the rasters derived by ArcGIS tools at higher window sizes $w = 30$ and 63 in Figure 5.3b. It was observed that the curvature results obtained for window size $w = 63$ appeared pixellated when compared to the similar window scale $w = 64$ in Figure 5.3a. This was because, the window scale $w = 63$ comprised of the original DEM that had been resampled to a factor of 21. The initial raster which had a resolution of 5m for each pixel values now had been converted to a raster where each pixel represented 105m on the ground. To elaborate this difference between the generated results, a side by side comparison of the NDVI is shown in Figure 5.4. The results obtained from the window-based aggregation (bottom row) entailed a detailed depiction as opposed to the conventional output (top row) which was pixellated.



(a) NDVI before applying Jenks Natural Breaks Algorithm

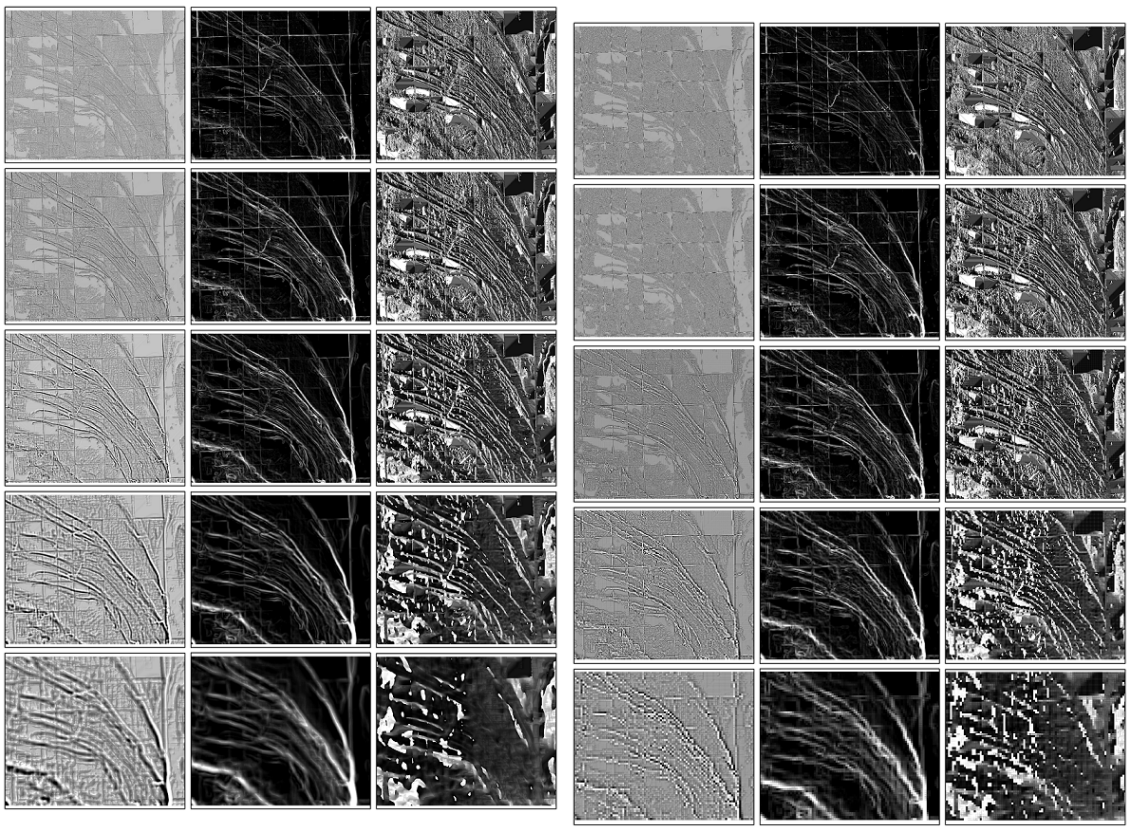
(b) NDVI of the study after applying Jenks Natural Breaks Algorithm

Figure 5.2. NDVI visualization of the study area after processing.

5.3.3. Traditional approach using ArcGIS

To obtain raster datasets in ArcGIS which were comparable to the one's generated using window-based aggregation, the DEM was resampled to a lower resolution first before proceeding to a larger window size. This process was done because the slope, curvature and aspect tools

of ArcGIS used a fixed window size of 3-by-3 for analysis. For the 3-by-3 window, that was comparable to the 4-by-4 from the proposed method, ArcGIS tools were run on the original DEM. To perform analysis comparable to 8-by-8, 16-by-16, 32-by-32 and 64-by-64 window sizes of the proposed method, the original DEM was resampled by using a window scale of 3-by-3, 5-by-5, 10-by-10 and 21-by-21 respectively. This was followed by running the 3-by-3 ArcGIS tools on the new DEMs. The output of raster datasets corresponding to curvature, aspect and slope now had window scales 9-by-9, 15-by-15, 30-by-30 and 63-by-63 respectively. The curvature, slope and aspect rasters obtained using ArcGIS is shown in Figure 5.3b.



(a) Sliding window aggregation output for window sizes 4, 8, 16, 32, and 64 represented from top-bottom

(b) Results obtained from ArcGIS for window sizes 3, 9, 15, 30 and 63 represented from top-bottom

Figure 5.3. DEM obtained from both methods. The GIS attributes shown include Curvature, Slope and Aspect from left-right.

5.4. Results

5.4.1. Random Forest based predictive modeling

To test the applicability of the proposed method in NDVI prediction using the derived landform attributes, Random Forest models were built for all window scales. A Random Forest model [97] operates by creating a set of decision trees based on the training data provided in the study. It uses a set of predictor variables and an outcome variable to develop a prediction model. Since the Random Forest model utilizes multiple decision trees to come to a conclusion, it addresses the over-fitting nature of a single decision tree algorithm [116, 117]. Such models usually create the entire decision tree without pruning.

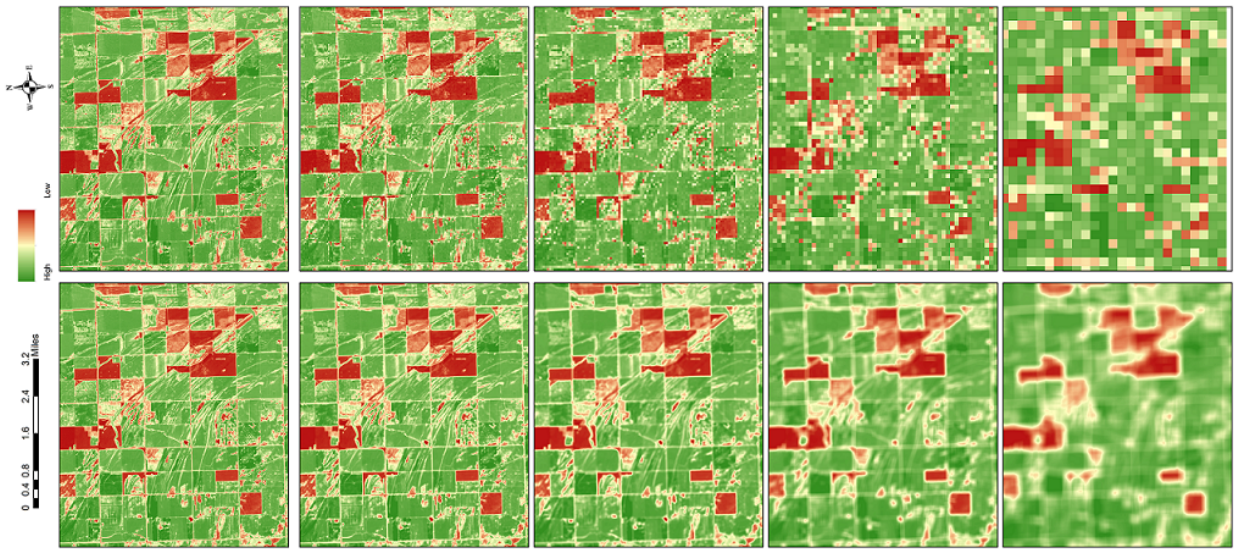


Figure 5.4. Results from window-based aggregation on NDVI (bottom row) where w = 4, 8, 16, 32 and 64 (left to right). Results from ArcGIS (top row) for w = 3, 9, 15, 30 and 63 (left to right).

Pruning has several advantages, one being that, it reduces over-fitting [118, 119] and the model can be applied to a vast array of testing data. It is also efficient as the entire tree does not have to be generated. However, pruning also reduces the accuracy of the model as it aims for a global solution. Random Forest models are based on the assumption that the prediction error rate

decreases by increasing the number of instances used for prediction [117]. A combination of N trees is used along with a selection of features that determines the best split. A split in the decision tree based on single or multiple predictor variables is considered best among all if it produces a node with high purity [120, 121]. A node is 100% pure if the split has all the records belonging to a single class [121]. Once all the trees are complete, the Random Forest model classifies each record in the dataset based on aggregate results obtained from N trees.

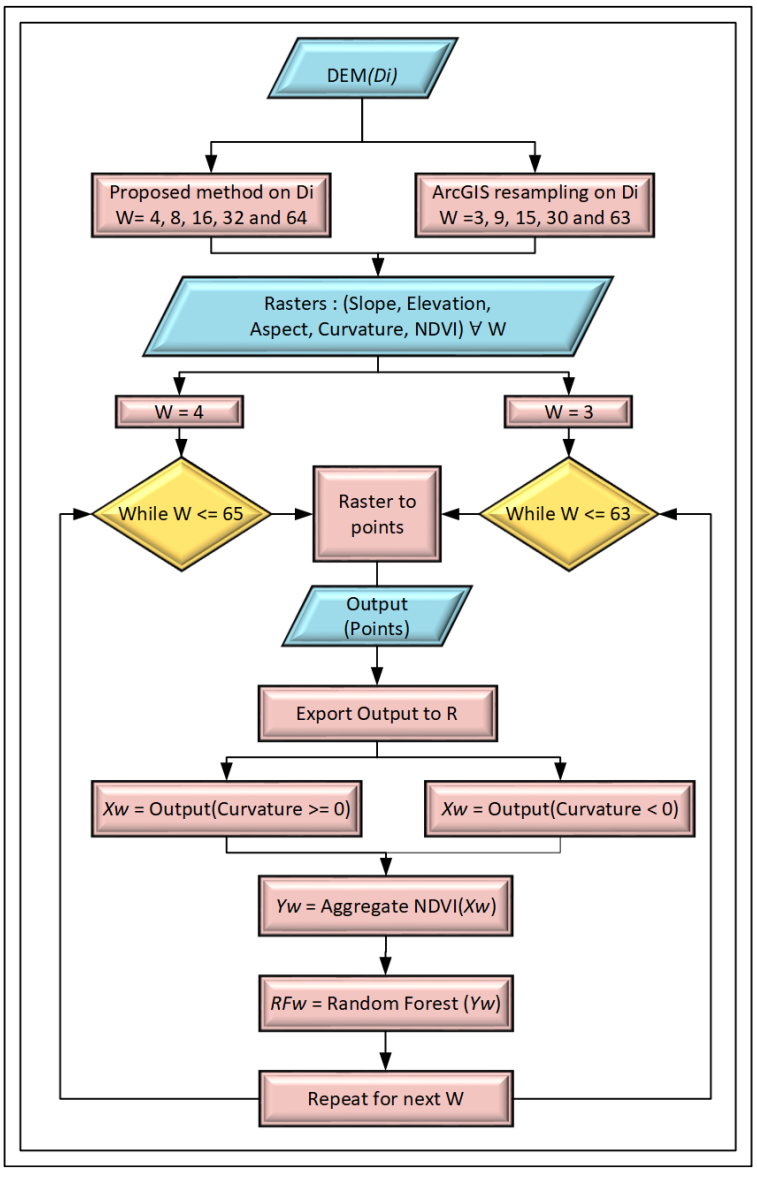


Figure 5.5. Steps for predictive modeling using Random Forest regression.

Table 5.1. Accuracy and GINI values generated from Random Forest regression models for both approaches.

Positive Curvature Results												
# of	Proposed method						ArcGIS method					
trees	Win size	OOB acc %	GINI impurity decrease				Win size	OOB acc %	GINI impurity decrease			
			Curv	Slope	Aspect	Elev			Curv	Slope	Aspect	Elev
500	4	94.67	34.75	29.34	6.6	17.05	3	31.96	16.04	17.86	13.63	23.09
	8	92.74	40.23	20.51	9.06	14.14	9	19.11	11.79	12.5	11.23	14.79
	16	84.09	32.34	18.61	14.45	15.4	15	14.28	6.34	6.67	6.18	7.71
	32	82.64	26.25	23.13	15.29	13.41	30	5.56	2.01	2.06	2.09	2.57
	64	92.09	2	1.96	1.04	1.51	63	-0.36	0.49	0.49	0.46	0.51
Negative Curvature Results												
500	4	84.04	37.97	23.84	8.26	16.24	3	51.89	19.82	19.87	10.38	23.2
	8	95.31	42.74	26.16	5.19	10.32	9	21.26	14.15	12.57	11.57	16.64
	16	84.46	30.74	24.13	8.12	17.85	15	10.5	7.64	7.44	7.76	9.32
	32	88.65	18.12	26.26	9.81	23.67	30	6.44	2.55	2.56	2.53	3.16
	64	94.5	19.33	23.72	6.97	14.67	63	-10.58	0.6	0.59	0.58	0.57

In this study, the landform attributes elevation, slope, aspect and curvature were used as predictor variables. The Random Forest model determined which variables were significant (produces the most efficient split) in NDVI prediction. The split was evaluated using Gini index [100, 122]. Five models were built for each of the window sizes $w = 4, 8, 16, 32$, and 64 obtained from the sliding window-based aggregation technique. This was followed by another five models built from the results generated by ArcGIS using window sizes $w = 3, 9, 15, 30$, and 63. The work-flow is shown in Figure 5.5. The input corresponded to the raster datasets for NDVI and all landform attributes derived from the DEM. The resultant table had curvature, aspect, slope, elevation and NDVI as attributes corresponding to one window size. The dataset was filtered according to the positive and negative curvature values to create two separate instances for the same study area. This was followed by an NDVI-based aggregation. The split based on curvature values was implemented due to the effect of an NDVI-based aggregation on the results. For example, if two areas with negative and positive curvature values having similar NDVI were to be aggregated, it would

produce erroneous results. Finally, a Random Forest model was implemented on the aggregates using the Random Forest package [42] in R. Since the attributes contained continuous values, the Random Forest package built a regression model using five hundred decision trees. This process was repeated for all the window sizes to obtain multiple Random Forest models. Results obtained for the positive and negative curvatures are summarized in Table 5.1.

5.4.2. Error analysis

Since the accuracy of the DEM played an integral role in the performance of every model and landform attributes derived from them; an error analysis was performed [123]. The difference in DEM values obtained from various window sizes were compared to the original 5m resolution DEM. The analysis was conducted on a randomly selected set of eight hundred points across the study area using equation 5.5 [124]. Here, x_i was chosen as the z-value of the original DEM and $x_i^{'}$ was the z-value of the DEM corresponding to larger window sizes which were obtained from ArcGIS and sliding window-based aggregation. Results are summarized in Figure 5.6.

$$RMSE = \sqrt{\frac{\sum (x_i - x_i{\prime})^2}{d}} \tag{5.5}$$

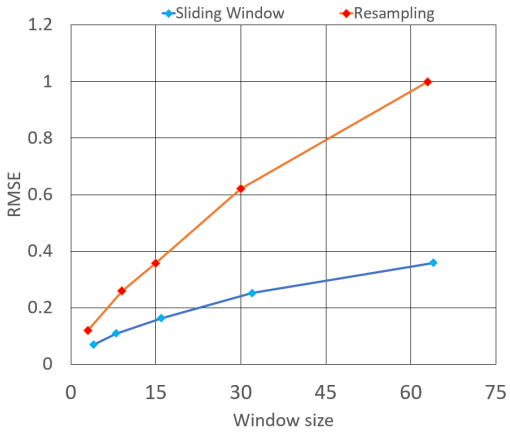


Figure 5.6. The graph shows a comparison of how RMSE increases with higher window sizes for the DEM.

5.4.3. Partial dependence plots

On identifying curvature and slope as the contributing variables in the study, their relation with the response variable NDVI was visualized using partial dependence plots [44]. Figure 5.7 explains the relation of slope and positive curvature with NDVI for the results generated from the window-based aggregation. The first row shows curvature variation with increasing window size of $w = 4$ and 8. The second row shows variation with window sizes 16 and 32 (left to right). The next row starts with $w = 64$ for curvature followed by the variation of NDVI with slope for $w = 4$. The fourth row shows variation of slope with $w = 8$ and 16. The fifth row shows slope variation with NDVI for $w = 32$ and 64. Figure 5.9 explains the relation of slope and negative curvature with NDVI for results generated from window-based aggregation. Partial dependence plots were also generated from results in ArcGIS across comparable window sizes $w = 3, 9, 15, 30$ and 63 as shown in Figure 5.8 for positive curvature and Figure 5.10 for negative curvature. All these figures follow a similar representation sequence as discussed for Figure 5.7

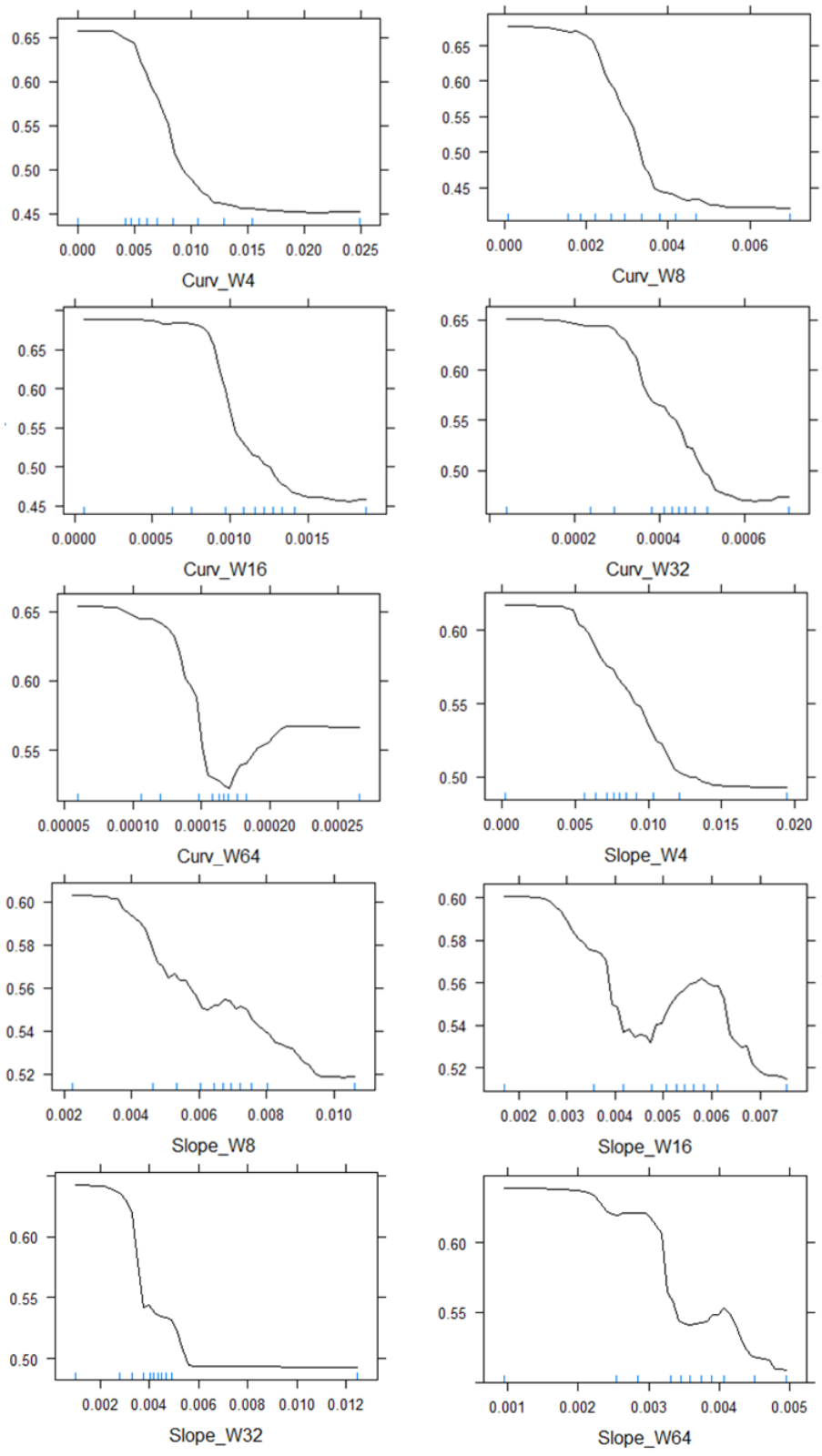


Figure 5.7. Partial dependence plots obtained for positive Curvature and Slope with NDVI for sliding window-based aggregation. Y-axis represents NDVI values and X-axis represents the slope or curvature value for the corresponding window size.

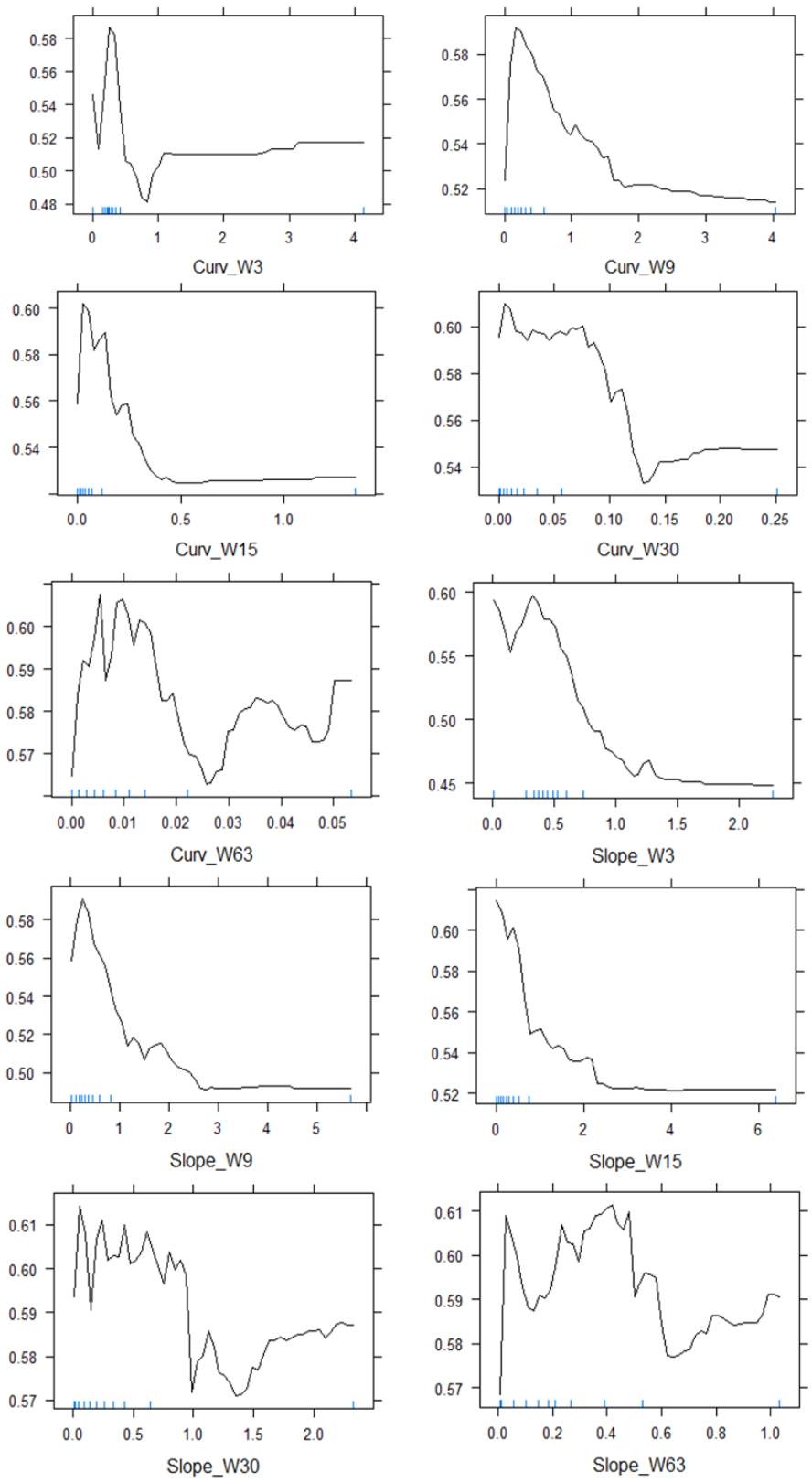


Figure 5.8. Partial dependence plots obtained for positive Curvature and Slope with NDVI for raster data derived from ArcGIS. Y-axis represents NDVI values and X-axis represents the slope or curvature value for the corresponding window size.

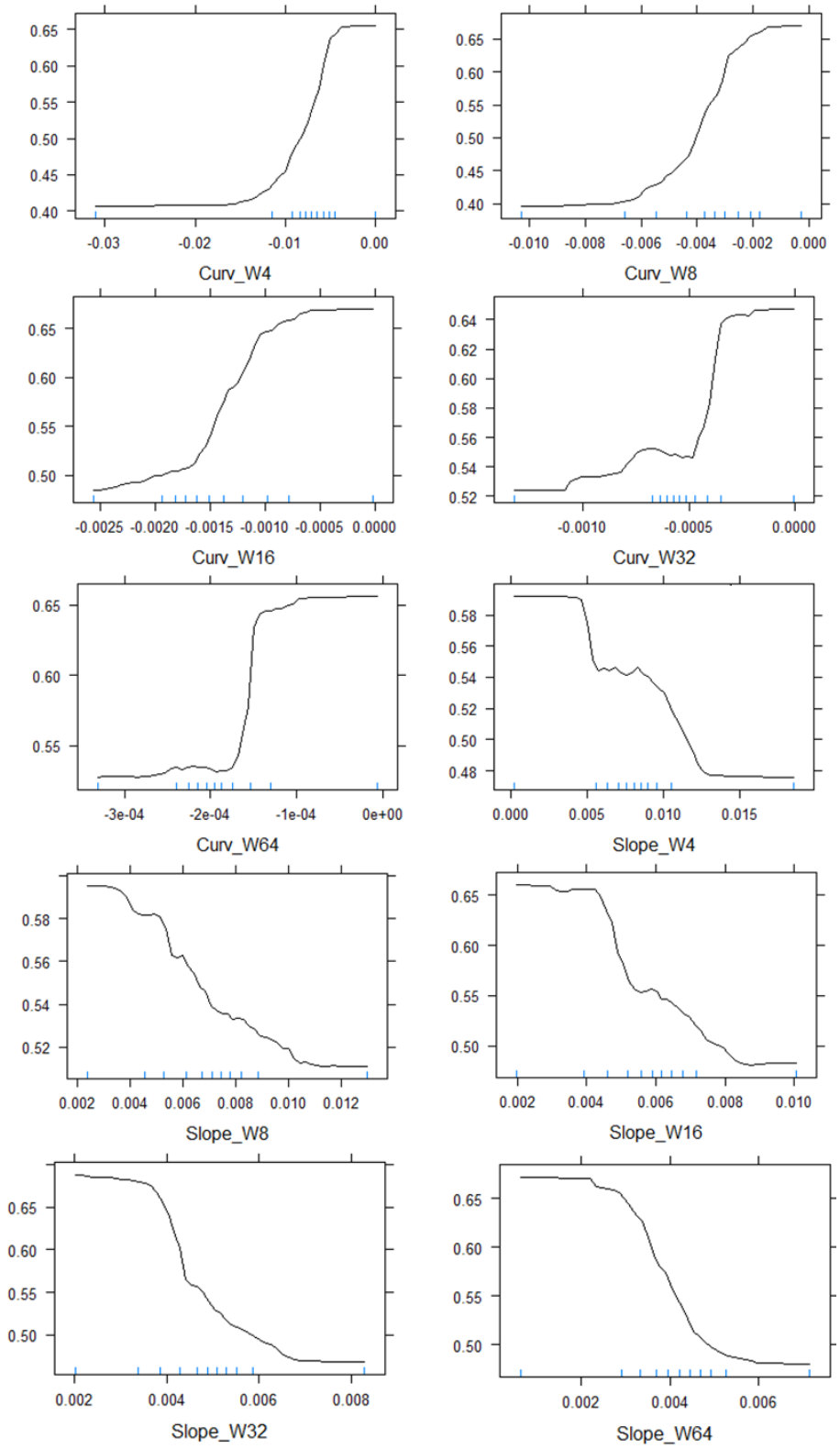


Figure 5.9. Partial dependence plots obtained for negative curvature and slope with NDVI for sliding window-based aggregation.Y-axis represents NDVI values and X-axis represents the slope or curvature value for the corresponding window size.

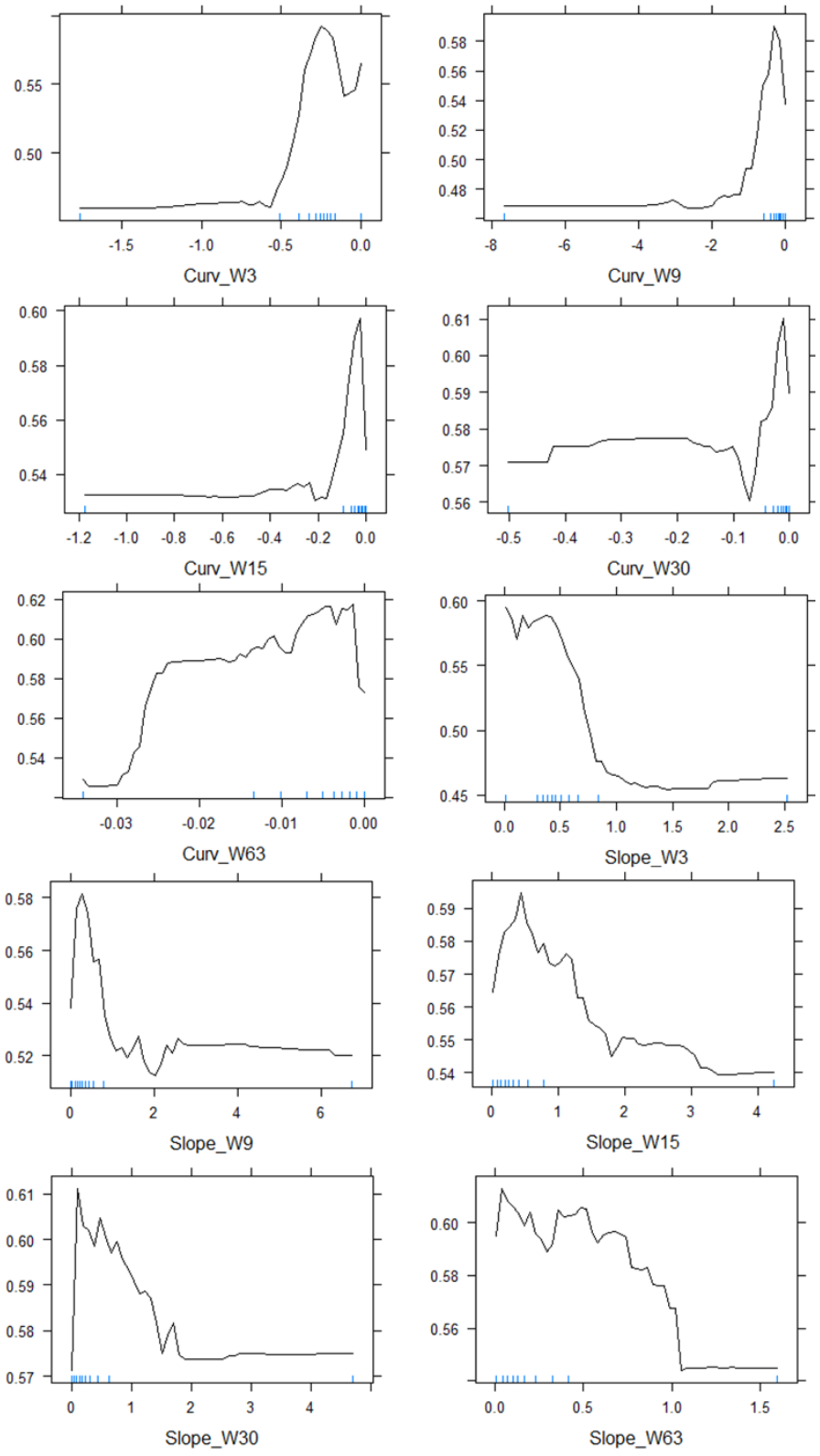
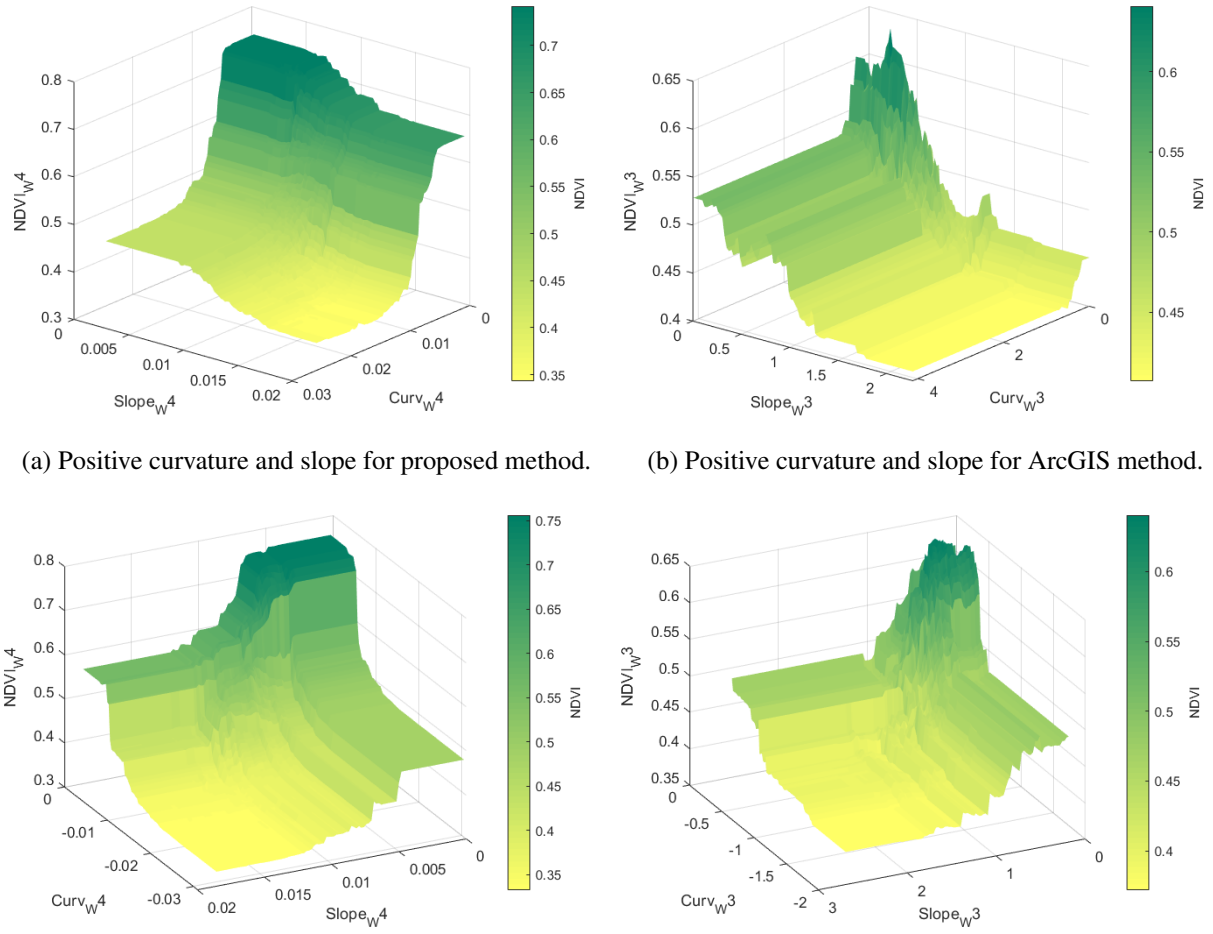


Figure 5.10. Partial dependence plots obtained for negative curvature and slope with NDVI for raster data derived from ArcGIS.Y-axis represents NDVI values and X-axis represents the slope or curvature value for the corresponding window size.

Figures 5.11a and 5.11b for window sizes 4 and 3 respectively shows a 3-D visualization of how NDVI varies when both positive curvature and slope are used as multi-predictors for sliding window-based aggregation and ArcGIS approach respectively. Similarly, Figures 5.11c and 5.11d show how NDVI varies with negative curvature and slope for the two approaches mentioned above.



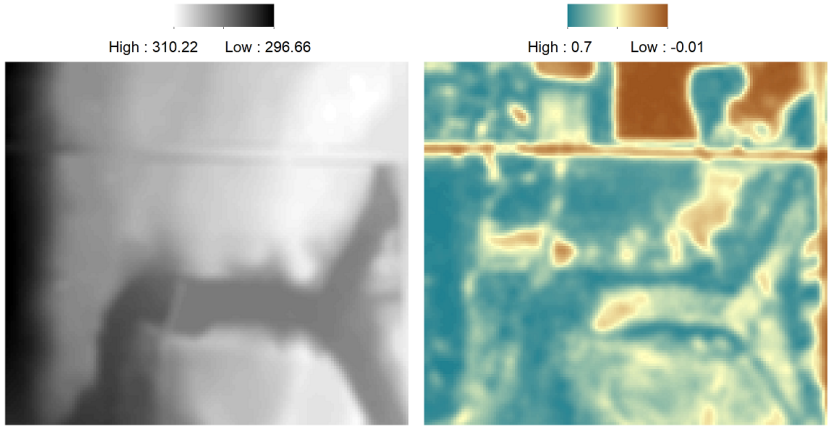
(a) Positive curvature and slope for proposed method.

(b) Positive curvature and slope for ArcGIS method.

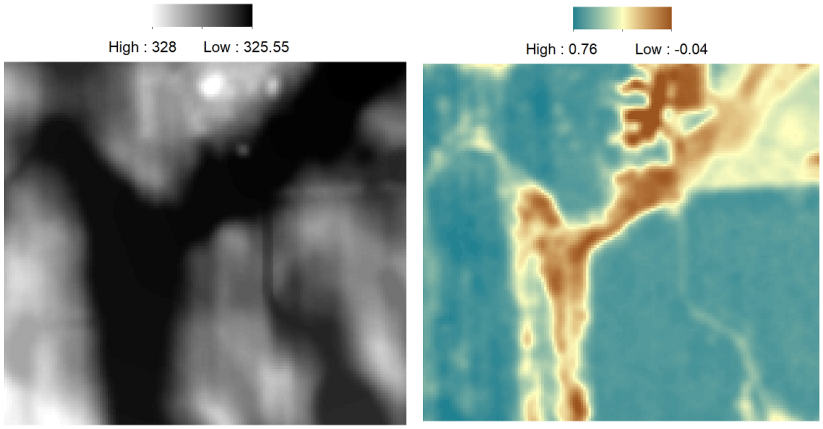
(c) Negative curvature and slope for proposed method.

(d) Negative curvature and slope for ArcGIS method.

Figure 5.11. 3-D multi-attribute partial dependence plots for Curvature and Slope with NDVI.



(a) Study area 1—Elevation (b) Study area 1—NDVI



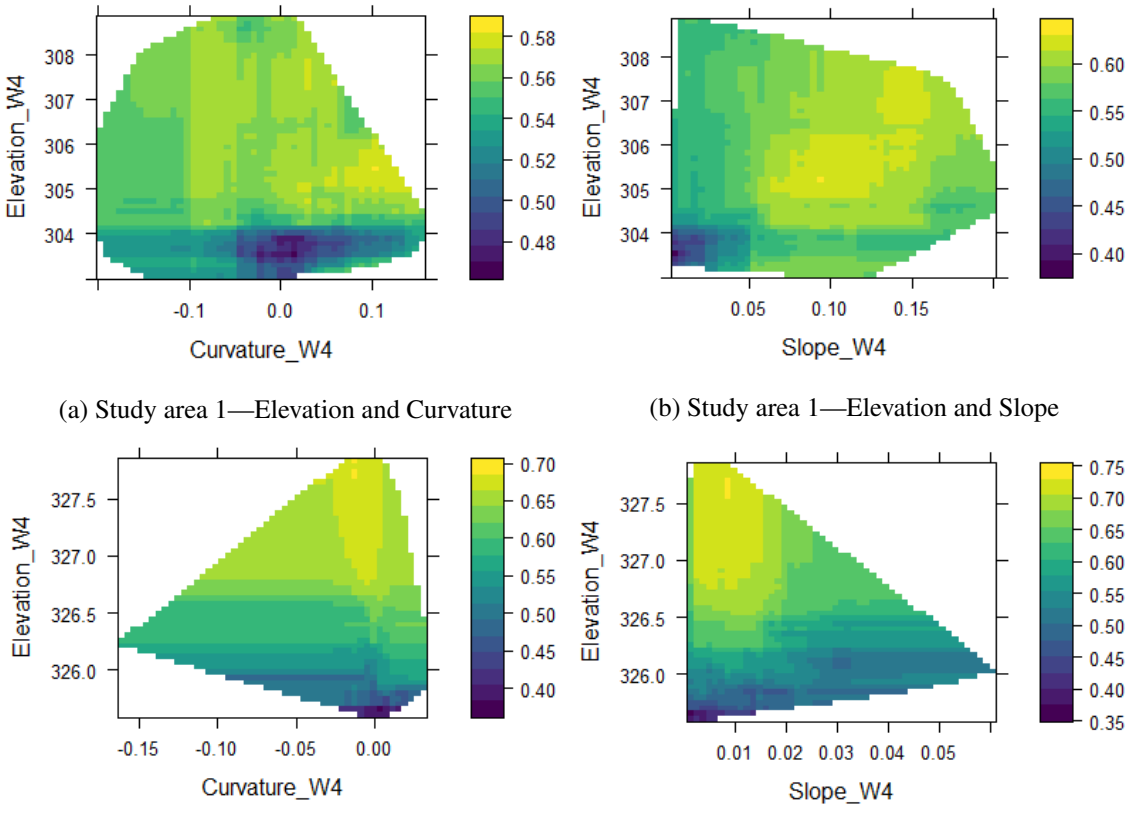
(c) Study area 2—Elevation (d) Study area 2—NDVI

Figure 5.12. DEM and NDVI values from depression study areas.

5.4.4. NDVI pattern in areas of depression

In this section, the window-based algorithm was applied on results obtained from a 4-by-4 window output to study how a localized depression on a field could effect yield. Depressions were classified as areas that had a relatively low elevation compared to the surrounding coupled with negligible slope and neutral curvature values. Two fields in the Richland county of North Dakota were considered for this study. These fields were relatively smaller in size compared to the previous evaluation of a larger area comprised of multiple fields. Figures 5.12a and 5.12b shows the elevation and the NDVI of the first study area respectively while Figures 5.12c and 5.12d shows the

elevation and the NDVI of the second study area respectively. Both of these areas have a depression which can be observed in their DEMs where the darker shade represents a lower elevation while the lighter shade corresponds to a higher elevation. In Figures 5.12b and 5.12d green represents a high while yellow represents a low NDVI value. Regions with shades of brown mostly represent barren lands having NDVI closer to zero. Figure 5.12b shows a large portion on the upper right corner which may be barren. This patch was ignored during the study. The NDVI and elevation along with curvature and slope corresponding to these areas were used to build Random Forest regression models followed by partial dependence plots which can be observed in Figure 5.13.



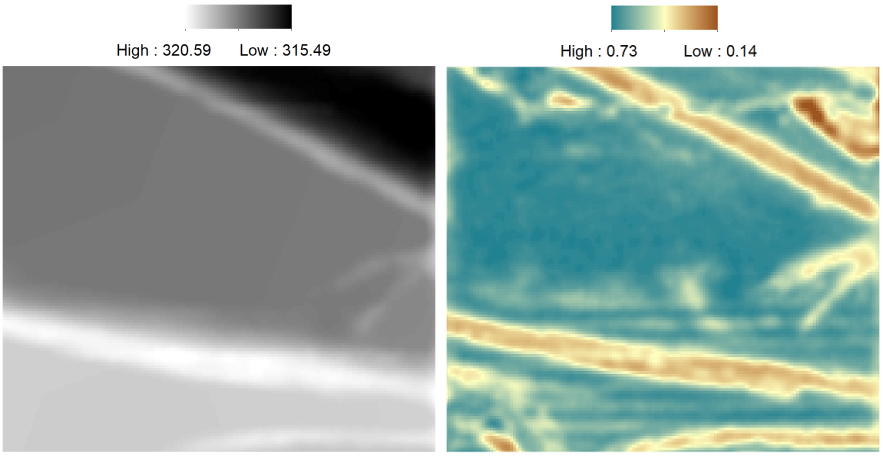
(a) Study area 1—Elevation and Curvature

(b) Study area 1—Elevation and Slope

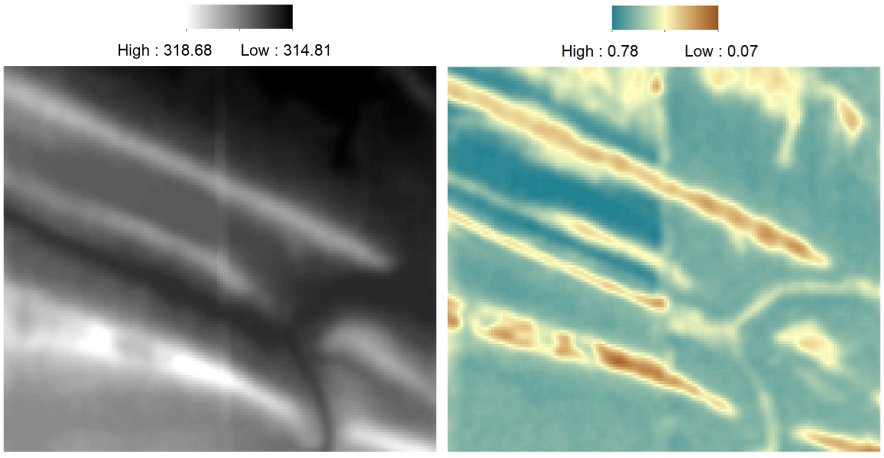
(c) Study area 2—Elevation and Curvature

(d) Study area 2—Elevation and Slope

Figure 5.13. Partial dependence plots (PDPs) corresponding to depression study areas with respect to NDVI.



(a) Study area 3—Elevation (b) Study area 3—NDVI



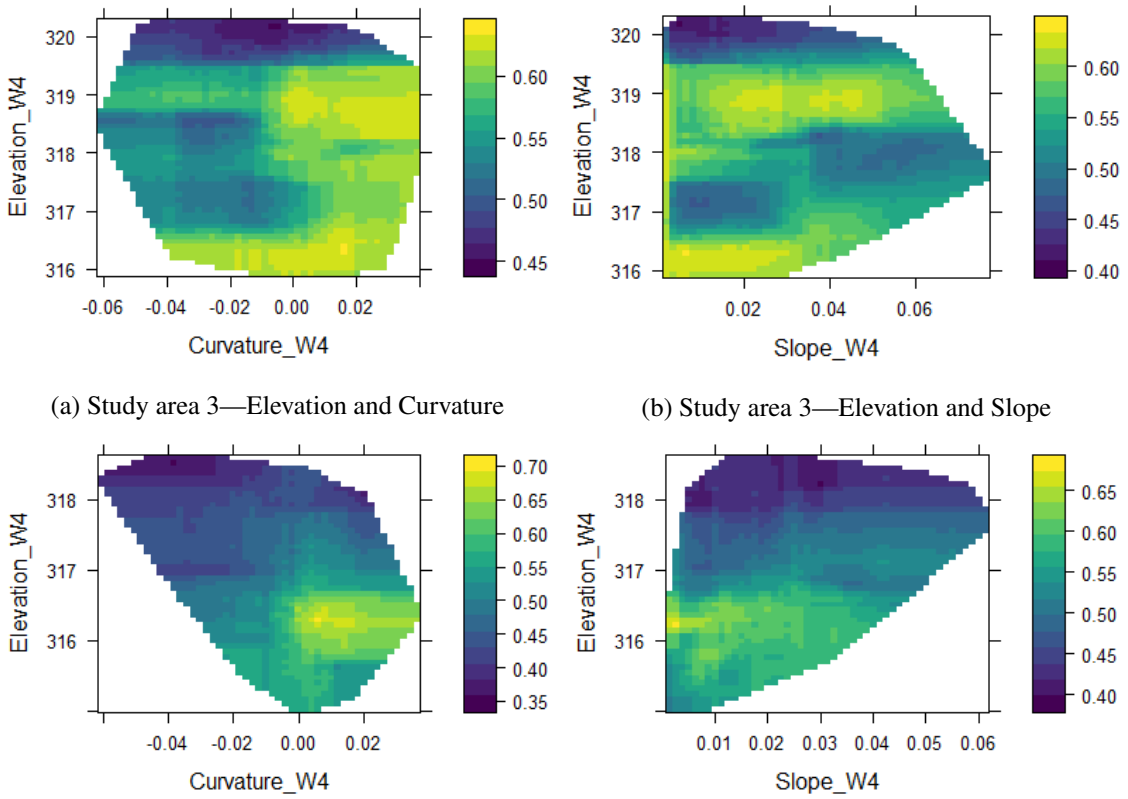
(c) Study area 4—Elevation (d) Study area 4—NDVI

Figure 5.14. DEM and NDVI values from elevated study areas.

5.4.5. NDVI pattern in highlands

In this section the algorithm was applied to areas that have a higher elevation compared to the surrounding areas. These areas mostly represent the top of localized hills that had a convex curvature. Two such study areas were used. Their NDVI and elevation values are shown in Figure 5.14. Both of the study areas also showed some regions of depression corresponding to the darker shade in the image. The higher grounds had a lighter shade and can be seen in Figures 5.14a and 5.14c. The elevation data in Figures 5.14c shows a patch of depression between two higher

grounds. The NDVI in both of these features were relatively low compared to the surrounding area as shown in Figure 5.14d. Random Forest models were also implemented for these areas where slope, curvature, and elevation values were used for predicting NDVI followed by generating partial dependence plots corresponding to the study areas. Results are shown in Figure 5.15 for study areas 3 and 4.



(a) Study area 3—Elevation and Curvature

(b) Study area 3—Elevation and Slope

(c) Study area 4—Elevation and Curvature

(d) Study area 4—Elevation and Slope

Figure 5.15. PDPs corresponding to elevated study areas with respect to NDVI.

5.5. Discussion

5.5.1. Random Forest based predictive modeling

5.5.1.1. Proposed method

R^2 values obtained from any prediction model have been successfully used for model comparison in the remote sensing domain [125, 126]. An analysis of the sliding window-based aggregation produced promising results across all the window scales. The R^2 values were estimated using the Out of Bag (OOB) accuracy for Random Forest models in $w = 4, 8, 16, 32$, and 64 windows sizes. The results in Table 5.1 showed that the positive curvature models could explain 94.67% , 92.74%, 84.09%, 82.64% and 92.09% of the NDVI values in the datasets respectively. The OOB accuracy was also evaluated for the models generated from the negative curvature. Here, the accuracy of the models were 84.04%, 95.31%, 84.46%, 88.65% and 94.5% respectively. The Gini index was calculated as the mean decrease in node impurity. A higher index signified that the attribute played a better role compared to other attributes in decreasing the node impurity thereby achieving a successful split. In all window sizes, the curvature attribute generated results with a higher node purity to create the Random Forest model followed by the slope. Attributes elevation and aspect varied across all the models with less significant contribution as shown in Table 5.1. Since the study area comprised of 1949-by-1604 pixels, which was relatively smaller (30.17 sq. miles.) than the two counties in which they belong, a window scale of 4 or 8 generated a model with the most relevant accuracy. A window size of $w = 64$ can be a potential solution to study yield across an entire county or state as it blends in deviations in the DEM which span fewer pixels. Although such a higher window scale has a generalization effect, it removes any DEM errors that could generate incorrect results making its accuracy comparable to lower window sizes.

All the window scales achieved high accuracies in NDVI prediction making the usage of sliding window-based aggregation justifiable.

5.5.1.2. Traditional method using ArcGIS

The results obtained from the ArcGIS resampling technique of reducing resolution did not show high accuracy like the proposed model. The Random Forest models for window sizes $w = 3, 9, 15, 30$, and 63 could explain only 31.96%, 19.11%, 14.28%, 5.56% and -0.36% of the variance respectively for the positive curvature and 51.89%, 21.26%, 10.5%, 6.44% and -10.58% respectively for the negative curvature. These values reflected that the conventional technique was incapable of finding any relationship of NDVI with the DEM derived attributes. Unlike the window-based aggregation where the curvature and slope significantly contributed to higher degree of node purity, the ArcGIS technique found elevation as the highest contributor to node purity as shown in Table 5.1. All the other attributes provided similar contribution and it was difficult to single out a DEM derived attribute that was better than the other. The model's accuracies decreases with increasing window size which can be expected based on the pixellated DEMs generated earlier in ArcGIS as shown in Figure 5.4. The window scale $w = 64$ produced a negative value for both curvature datasets. This negative value is usually encountered in R if the numerator shown in equation 5.6 [127] which represents the mean squared error is larger than the variance of y values in the denominator [127, 116]. Here y and y_i represents the observed and predicted values respectively. $\bar{y}$ represents the mean of all the observation. The model generated cannot be used to derive any significant relationship among the attributes. Thus, the conventional technique was not suitable for predictive analysis over multiple window scales in its present form.

$$R^2 = 1 - \frac{\frac{1}{n}\sum(y - y_i)^2}{\frac{1}{n}\sum(y - \bar{y})^2} \tag{5.6}$$

5.5.2. Error analysis

The increase in variance of the DEM across multiple scales from their original values is shown in Figure 5.6. RMSE of the window-based aggregation method was considerably lower when compared to the ArcGIS methodology. The highest RMSE at window size $w = 64$ was almost three times less than the one's obtained at the window size $w = 63$ using ArcGIS based approach. The rate of increase in RMSE was also lower for window-based aggregation when compared to its counterpart showing that the elevation data loses less information over higher window sizes in the sliding window-based aggregation.

5.5.3. Partial dependence plots

Since slope and curvature provided the highest contribution to Random Forest models, partial dependence plots were generated for NDVI with respect to these two attributes. In Figure 5.7, a smooth transition was observed from high to low NDVI with an increase in both slope and positive curvature values showing that NDVI was highest in regions which had minimal undulations in the soil surface. These results can be justified since areas with high slope and curvature values could also be associated with undulations. These undulations may show low yield due to factors such as running water causing surface erosion. A similar pattern was also observed in the negative curvature plots shown in Figure 5.9. These plots showed that NDVI decreased with an increase in the negative curvature values and an increase in the slope as well. The results also concur with the idea that areas with negative curvature and high slope mostly represented an undulating surface that may be unproductive.

Partial dependence plots derived from results processed in ArcGIS further explained the low accuracy rates generated by the Random Forest models. A closer look at windows $w = 3$ and 9 for curvature-NDVI plots in Figure 5.8 showed that the NDVI increased with an increase in

curvature first before dropping. This irregularity propagated at higher window scales, rendering the model generated by window scale $w = 63$ completely unfit for any further analysis.

For positive curvature, both the models were able to show similar patterns as was evident from the sudden increase in NDVI mid-way across the curvature plots for window scales $w = 63$ in Figure 5.8 and window scale $w = 64$ in Figure 5.7. This pattern, which was invisible at a lower window scales and appeared at the larger one was a source of interest. On further analysis it was verified that the positive curvature was also associated with the roads because they had a higher elevation and a convex shape like the elevated areas. It appeared that at a higher resolution, the roads would increase in width and spread across a larger window due to the averaging effect that occurred during window-based aggregation. This would cause the curvature values to increase and span across a larger area in the image. A similar averaging effect was visible in the NDVI values but with a different outcome. The regions where the roads were present in the image had a very low NDVI while fields adjacent to it had a comparatively higher NDVI. Both these values yielded an average NDVI on aggregation. So now, we have roads with positive curvature values spanning a large window which overlapped with average NDVI values spanning the same window. This effect was visible in the graph as a sudden increase in the NDVI values with the positive curvature. This theory was further validated by the fact that the irregularity was not observed in the negative curvature graphs because the negative curvature would not consider roads in the first place. This effect was also not observed at lower window scales because the roads were usually restricted to fewer pixels. Even if both aggregation and traditional approaches were able to detect this pattern, the sliding window-based method did better job as the DEMs can hold more information. The results generated in ArcGIS did not show a gradual change in pattern due to pixellation.

In Figure 5.11a, which shows the NDVI variation with slope and curvature as multi-predictors, a smooth transition between adjacent pixels were observed and a noticeable trend generated using the sliding window-based aggregation. The 3-D plot generated by the conventional method for window scale $w = 3$ in Figure 5.11b had irregular peaks due to the high amount of pixellation in the dataset caused by resampling to a higher resolution. The 3-D plots corresponding to the negative curvature in Figure 5.11c and 5.11d showed a similar trend of irregularity in the ArcGIS output compared to the window-based aggregation. The NDVI was minimum at high negative curvature and low slope. The sliding window-based aggregation did a better job of smoothing the effect of pixellation that caused the irregular results in the ArcGIS output.

5.5.4. NDVI pattern in areas of depression

Two study areas which had a relatively low elevation compared to their surroundings were also evaluated as shown in Figures 5.12a and 5.12c. As discussed earlier, these regions represented a single field compared to our previous Random Forest models and partial dependence plots which represented a larger area. The NDVI for these regions is also shown in Figures 5.12b and 5.12d respectively. After constructing the Random Forest models, partial dependence plots for these areas were derived as shown in Figure 5.13. Figures 5.13a and 5.13c shows how the NDVI varied with elevation and curvature for these two study areas. It was observed that the NDVI was lowest in areas that had moderate to low curvature and low elevation. Figures 5.13b and 5.13d shows the variation of NDVI with respect to the slope and elevation. It was observed that the NDVI was lowest in areas of negligible slope and low elevation. These results showed that a depression in the land usually had lower NDVI values compared to its surrounding area. This may be due to water-logging which reduces the crop yield [128]. Unlike the plots generated on a macro-scale which showed that the NDVI decreased with an increase in both positive and negative curvature,

these results also showed that the NDVI could be lower in minimum curvature and slope values as well. Areas of depression were likely to suffer from a water stress that could potentially cause the crop yield to decrease. These regions mostly consisted of negative curvature values showing that the land was concave. It should be noted that an effort was made to classify the depression using the ArcGIS method. The DEM derived attributes were highly pixellated and did not yield any noticeable relationship at the scale comparable to a single field. This may be one of the reasons as to why depression was not readily visible on a macro-level and they often got ignored during the study to find the overall trend across a farmland. The proposed algorithm not only succeeded on a macro-scale but can also show the results otherwise obscured by the problem of low resolution in conventional methods.

5.5.5. NDVI pattern in highlands

Figures 5.14a and 5.14c also showed regions at the scale of a single field which had a higher elevation with respect to its surrounding. It was observed in Figures 5.14b and 5.14d that the NDVI for these regions were comparatively lower. Just like in depressions, partial dependence plots for these areas were generated as shown in Figure 5.15. Results showed that the NDVI remained low across the entire curvature range for the higher elevation. This was due to the heterogeneity of the study area. Both of these places had a mix of elevated areas and some patches of depression. The depression was more visible in the study area 4 as shown in Figure 5.14c. A similar pattern was also visible in Figures 5.15b and 5.15d where the elevation and slope were compared together with the NDVI. Higher elevation values with negligible slope mostly had the lowest NDVI. From these results it can be concluded that the NDVI decreased when the elevation of a place was comparatively higher than its surrounding. The higher elevation may have affected the optimum water-table depth [129] required for good yield. Both the study areas demonstrated the importance of optimum

water-table depth at a local scale on a field for maximum crop yield. Any change in elevation, be it higher or lower than the surrounding values, had a potential of impacting the quality of crops.

6. CONCLUSIONS

Since the advent of geospatial data mining, GIS algorithms were developed around spatial datasets with a resolution close to 60m [35]. The resolution, however, has increased dramatically with new passive sensors. Lidar has also enabled accurate mapping with point cloud spacing of 1m [37]. With such high resolution images, a fixed 3-by-3 cell analysis, may not be a feasible solution. In this dissertation, a multiscalar sliding window-based aggregation technique was implemented to derive landform attributes slope, aspect and curvature to address the shortcomings of existing GIS algorithms.

The proposed aggregation methodology was used to derive slope, aspect and curvature on several scales, where results from one scale where reused in the next scaling process. Reusing results allowed the computation time to be logarithmic and the output obtained from the proposed aggregation were much less subject to noise when compared to the conventional approach.

Error analysis was conducted on results derived by the proposed method using window scales $w = 4, 8, 16, 32$ and 64 which were compared with the error propagation for the traditional approach of comparable window scales $w = 3, 9, 15, 30$ and 63. The proposed aggregation was subject to less error propagation compared to the conventional approach. For the proposed aggregation, the minimum and maximum RMSE recorded were 0.102 and 0.124 for $w = 64$ and $w = 32$ respectively as shown in Table 3.1. However, the minimum and maximum RMSE recorded for conventional method were 0.179 and 0.473 for $w = 3$ and $w = 63$ respectively. For the proposed aggregation, the RMSE values were very close to each other with a mean of 0.1076 and a standard deviation of 0.0135. The resampling approach showed a constant increase in RMSE values with the increasing length scale. The mean RMSE was 0.371 and the standard deviation was 0.105

which was almost ten times more than the standard deviation of the proposed aggregation. Visual interpretation of the derived results showed that the proposed model retained patterns in the DEM even when upscaling was performed whereas the ArcGIS derived attributes mostly suffered from increased pixellation at larger window scales.

Predictive models from the results derived using the proposed approach also outperformed the conventional approach. Two classification models that were compared included Random Forest and Naive Bayes. Results indicated that the proposed method could generate models with a higher NDVI prediction accuracy and can be used to study relationship of DEM derived attributes with yield. It is worth mentioning that even though the accuracies derived from proposed method were more than ArcGIS results, the highest accuracy was achieved using Random Forest and not Naive Bayes classification. Random Forest produced 74.47% accuracy for window scale $w = 4$ on testing dataset as shown in Table 4.5. Also, the accuracy reduced only to 71.17% for window scale $w = 16$. In comparison, the accuracy for Naive Bayes classification was 73.57% at $w = 4$ and reduced to 67.21% at $w = 16$ as shown in Table 4.3.

The window-based aggregation was also used to produce Random Forest regression models and study the Gini index corresponding to attributes slope, aspect, curvature and elevation that were used to derive NDVI. These indexes shown in Table 5.1 consistently showed higher values for the curvature and slope attributes in model creation, indicating that they were the most significant attributes contributing to yield estimation. The regression models also achieved a higher accuracy of 95.31% in predicting NDVI for $w = 4$ compared to a 51.89% obtained for window size $w = 3$ in the traditional model. Predictive models were also derived from small regions of depressions and elevated areas in other DEMs. Both these features showed signs of low NDVI.

Finally a subset of results obtained from the proposed aggregation were used to generate simulated DEMs using semivariograms to test how much information was lost in the resampling process. IGF values were used to evaluate the simulated DEMs. It was observed that the simulated DEMs created using values from the proposed model had lower IGF values than the models generated using data from ArcGIS. Lower IGF values indicated that the results from proposed approach had a lower error rate in generating simulated DEMs compared to the conventional one. The exponential semivariogram model had an IGF value of 0.1045 for $w = 64$ in the proposed model compared to 0.3697 for $w = 63$ in the conventional approach as shown in Table 3.3. In both cases, increasing the length scales reduced the amount of topographic detail that can be interpolated.

In future work, the aggregation method would be applied across other areas to determine its consistency. SOM-based methodology [130] would be implemented to identify a different grouping scheme for categorizing the dataset to aid in classification. SOM or self-organizing maps use a clustering approach to find related groups in a dataset [130]. Since SOM uses a learning rate based on the data, it can adjust the grouping with emphasis on the dataset without following a set of predetermined rules like the natural breaks or quantile. The user has to input factors such as the number of clusters and the learning rate for the system to obtain the results. The result variation with floating point accuracy and adjusting bit depth of the multispectral images would also be assessed. An attempt to compare resampling output with proposed aggregation would be done after application of interpolation methods such as spline [131] on the results derived from ArcGIS resizing. Since spline interpolation can generate a smooth surface [132], a comparison would reveal how much information is lost during this interpolation compared to the proposed sliding window-based aggregation.

REFERENCES

[1] Anne M Denton, Mostofa Ahsan, David Franzen, and John Nowatzki. Multi-scalar analysis of geospatial agricultural data for sustainability. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 2139–2146. IEEE, 2016.

[2] ESRI. What is gis ? `https://www.esri.com/en-us/what-is-gis/overview`, June 2019.

[3] Paul Bolstad. *GIS Fundamentals: A First Text on Geographic Information Systems.* Eider Press, 2005.

[4] John A Richards and JA Richards. *Remote sensing digital image analysis*, volume 3. Springer, 1999.

[5] Hans-Erik Andersen, Robert J McGaughey, and Stephen E Reutebuch. Estimating forest canopy fuel parameters using lidar data. *Remote sensing of Environment*, 94(4):441–449, 2005.

[6] Michael Hutchinson and J Gallant. Digital elevation models. *Terrain analysis: principles and applications*, pages 29–50, 2000.

[7] Pierre-Jean Lapray, Xingbo Wang, Jean-Baptiste Thomas, and Pierre Gouton. Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626–21659, 2014.

[8] Zahra Sadeghipoor, Jean-Baptiste Thomas, and Sabine Süsstrunk. Demultiplexing visible and near-infrared information in single-sensor multispectral imaging. In *Color and Imaging Conference*, volume 2016, pages 76–81. Society for Imaging Science and Technology, 2016.

[9] Douglas M Chabries, Steven W Booras, and Gregory H Bearman. Imaging the past: recent applications of multispectral imaging technology to deciphering manuscripts. *Antiquity*, 77(296):359–372, 2003.

[10] John B Adams, Donald E Sabol, Valerie Kapos, Raimundo Almeida Filho, Dar A Roberts, Milton O Smith, and Alan R Gillespie. Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the brazilian amazon. *Remote sensing of Environment*, 52(2):137–154, 1995.

[11] B Tyler Wilson, Andrew J Lister, and Rachel I Riemann. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. *Forest Ecology and Management*, 271:182–198, 2012.

[12] Papia F Rozario, Peter Oduor, Larry Kotchman, and Michael Kangas. Quantifying spatiotemporal change in landuse and land cover and accessing water quality: a case study of missouri watershed james sub-region, north dakota. *Journal of Geographic Information System*, 8(06):663, 2016.

[13] Papia F Rozario, Peter Oduor, Larry Kotchman, and Michael Kangas. Transition modeling of land-use dynamics in the pipestem creek, north dakota, usa. *Journal of Geoscience and Environment Protection*, 5(03):182, 2017.

[14] Yichun Xie, Zongyao Sha, and Mei Yu. Remote sensing imagery in vegetation mapping: a review. *Journal of plant ecology*, 1(1):9–23, 2008.

[15] Sebastian Candiago, Fabio Remondino, Michaela De Giglio, Marco Dubbini, and Mario Gattelli. Evaluating multispectral images and vegetation indices for precision farming applications from uav images. *Remote sensing*, 7(4):4026–4047, 2015.

[16] Hao Zhang, Lei Xi, Xinming Ma, Zhongmin Lu, Yali Ji, and Yanna Ren. Research and development of the information management system of agricultural science and technology to farmer based on gis. In *International Conference on Computer and Computing Technologies in Agriculture*, pages 141–150. Springer, 2007.

[17] Robert Hickey. Slope angle and slope length solutions for gis. *Cartography*, 29(1):1–8, 2000.

[18] Kang-tsung Chang and Bor-wen Tsai. The effect of dem resolution on slope and aspect mapping. *Cartography and geographic information systems*, 18(1):69–77, 1991.

[19] Sarah O Tweed, Marc Leblanc, John A Webb, and Maciek W Lubczynski. Remote sensing and gis for mapping groundwater recharge and discharge areas in salinity prone catchments, southeastern australia. *Hydrogeology Journal*, 15(1):75–96, 2007.

[20] ESRI. How slope works. `http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-slope-works.htm`, June 2019.

[21] ArcMap. How aspect works. `http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-aspect-works.htm`, June 2019.

[22] Ian S Evans. An integrated system of terrain analysis and slope mapping. *Zeitschrift fur Geomorphologie*, 36:274–295, 1980.

[23] Lucian Blaga. Aspects regarding the significance of the curvature types and values in the studies of geomorphometry assisted by GIS. *Annals of the University of Oradea, Geography Series/Analele Universitatii din Oradea, Seria Geografie*, 22(2), 2012.

[24] Peter A Burrough, Rachael McDonnell, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford university press, 2015.

[25] XM Jin, YK Zhang, ME Schaepman, JGPW Clevers, and Zhongbo Su. Impact of elevation and aspect on the spatial distribution of vegetation in the qilian mountain area with remote sensing data. *The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences*, 37:1385–1390, 2008.

[26] Yury Dvornikov, Artem Khomutov, Damir Mullanurov, Ksenia Ermokhina, Anatoly Gubarkov, and Marina Leibman. Gis and field data based modelling of snow water equivalent in shrub tundra. *Fennia*, 193(1):53–65, 2015.

[27] A Crave and C Gascuel-Odoux. The influence of topography on time and space distribution of soil surface water content. *Hydrological processes*, 11(2):203–210, 1997.

[28] Vincent Chaplot, Christian Walter, and Pierre Curmi. Improving soil hydromorphy prediction according to dem resolution and available pedological data. *Geoderma*, 97(3-4):405–422, 2000.

[29] RF Follett, EJ Doering, GA Reichman, and LC Benz. Effect of irrigation and water-table depth on crop yields 1. *Agronomy Journal*, 66(2):304–308, 1974.

[30] Estela Nadal-Romero, Kristien Petrlic, Els Verachtert, Esther Bochet, and Jean Poesen. Effects of slope angle and aspect on plant cover and species richness in a humid mediterranean badland. *Earth Surface Processes and Landforms*, 39(13):1705–1716, 2014.

[31] AS El-Hassanin, TM Labib, and EI Gaber. Effect of vegetation cover and land slope on runoff and soil losses from the watersheds of burundi. *Agriculture, ecosystems & environment*, 43(3-4):301–308, 1993.

[32] Sushil Pradhan. Crop area estimation using gis, remote sensing and area frame sampling. *International Journal of Applied Earth Observation and Geoinformation*, 3(1):86–92, 2001.

[33] Bunkei Matsushita, Wei Yang, Jin Chen, Yuyichi Onda, and Guoyu Qiu. Sensitivity of the enhanced vegetation index (evi) and normalized difference vegetation index (ndvi) to topographic effects: a case study in high-density cypress forest. *Sensors*, 7(11):2636–2651, 2007.

[34] Darrel L Williams, Samuel Goward, and Terry Arvidson. Landsat. *Photogrammetric Engineering & Remote Sensing*, 72(10):1171–1178, 2006.

[35] USGS. Mapping, remote sensing, and geospatial data. `https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?qt-news_science_products=0#qt-news_science_products`, July 2019.

[36] R Douglas Ramsey, Dennis L Wright Jr, and Chris McGinty. Evaluating the use of landsat 30m enhanced thematic mapper to monitor vegetation cover in shrub-steppe environments. *Geocarto International*, 19(2):39–47, 2004.

[37] Minnesota Geographic Metadata Guidelines. Lidar elevation, arrowhead region, ne minnesota, 2011. http://www.mngeo.state.mn.us/chouse/metadata/lidar_arrowhead2011.html, June 2019.

[38] Txomin Hermosilla, Luis A Ruiz, Jorge A Recio, and Javier Estornell. Evaluation of automatic building detection approaches combining high resolution images and lidar data. *Remote Sensing*, 3(6):1188–1210, 2011.

[39] Jon Atli Benediktsson, Jocelyn Chanussot, and Wooil M Moon. Very high-resolution remote sensing: Challenges and opportunities [point of view]. *Proceedings of the IEEE*, 100(6):1907–1910, 2012.

[40] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[41] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[42] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[43] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[44] Brandon M Greenwell. pdp: an r package for constructing partial dependence plots. *The R Journal*, 9(1):421–436, 2017.

[45] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[46] Mark A Friedl, Carla E Brodley, and Alan H Strahler. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):969–977, 1999.

[47] Kai Liu, Xia Li, Xun Shi, and Shugong Wang. Monitoring mangrove forest changes using remote sensing and gis data with decision-tree learning. *Wetlands*, 28(2):336, 2008.

[48] Rahul Gomes, Anne M Denton, and David Franzen. Comparing classification accuracy of ndvi with dem derived attributes using multi-scalar approach in geographic information systems. In *Electro Information Technology (EIT), 2019 IEEE International Conference on*. IEEE, 2019.

[49] Anne M Denton, Rahul Gomes, and David Franzen. Scaling up window-based slope computations for geographic information systems. In *Electro Information Technology (EIT), 2018 IEEE International Conference on*. IEEE, 2018.

[50] Rahul Gomes, Anne Denton, and David Franzen. Quantifying efficiency of sliding-window based aggregation technique by using predictive modeling on landform attributes derived from dem and ndvi. *ISPRS International Journal of Geo-Information*, 8(4):196, 2019.

[51] M Albani*, B Klinkenberg, DW Andison, and JP Kimmins. The choice of window size in approximating topographic surfaces from digital elevation models. *International Journal of Geographical Information Science*, 18(6):577–593, 2004.

[52] John Nikolaus Callow, Kimberly P Van Niel, and Guy S Boggs. How does modifying a dem to reflect known hydrology affect subsequent terrain analysis? *Journal of hydrology*, 332(1-2):30–39, 2007.

[53] Mingteh Chang. *Forest hydrology: an introduction to water and forests*. CRC press, 2006.

[54] Steven D Warren, Matthew G Hohmann, Karl Auerswald, and Helena Mitasova. An evaluation of methods to determine slope using digital elevation data. *Catena*, 58(3):215–233, 2004.

[55] Kenneth G Renard, George R Foster, Glenn A Weesies, and Jeffrey P Porter. Rusle: Revised universal soil loss equation. *Journal of soil and Water Conservation*, 46(1):30–33, 1991.

[56] M Irfan Ashraf, Zhengyong Zhao, Charles P-A Bourque, and Fan-Rui Meng. Gis-evaluation of two slope-calculation methods regarding their suitability in slope analysis using high-precision lidar digital elevation models. *Hydrological Processes*, 26(8):1119–1133, 2012.

[57] Stefan Kienzle. The effect of dem raster resolution on first order, second order and compound terrain derivatives. *Transactions in GIS*, 8(1):83–111, 2004.

[58] Don P Mitchell and Arun N Netravali. Reconstruction filters in computer-graphics. In *ACM Siggraph Computer Graphics*, volume 22, pages 221–228. ACM, 1988.

[59] Lubos Mitas and Helena Mitasova. Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, 1(2), 1999.

[60] Simon Wu, Jonathan Li, and GH Huang. A study on dem-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. *Applied Geography*, 28(3):210–223, 2008.

[61] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

[62] Matthew Dunn and Robert Hickey. The effect of slope algorithms on slope estimates within a gis. *Cartography*, 27(1):9–15, 1998.

[63] Michael R Travis, Gary H Elsner, Wayne D Iverson, and Christine G Johnson. Viewit: computation of seen areas, slope, and aspect for land-use planning. *Gen. Tech. Rep. PSW-GTR-11. Berkeley, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture: 70 p*, 11, 1975.

[64] Zhengyong Zhao, Glenn Benoy, Thien Lien Chow, Herb W Rees, Jean-Louis Daigle, and Fan-Rui Meng. Impacts of accuracy and resolution of conventional and lidar based dems on parameters used in hydrologic modeling. *Water resources management*, 24(7):1363–1380, 2010.

[65] R Srinivasan and BA Engel. Effect of slope prediction methods on slope and erosion estimates. *Applied Engineering in Agriculture*, 7(6):779–783, 1991.

[66] P Jiang and KD Thelen. Effect of soil and topographic properties on crop yield in a north-central corn–soybean cropping system. *Agronomy Journal*, 96(1):252–258, 2004.

[67] Jonathan Bennie, Mark O Hill, Robert Baxter, and Brian Huntley. Influence of slope and aspect on long-term vegetation change in british chalk grasslands. *Journal of ecology*, 94(2):355–368, 2006.

[68] Aspect (geography). Aspect in physical geography. `http://wiki.gis.com/wiki/index.php/Aspect_(geography)#cite_note-1`, June 2019.

[69] M Agassi, J Morin, and I Shainberg. Slope, aspect, and phosphogypsum effects on runoff and erosion. *Soil Science Society of America Journal*, 54(4):1102–1106, 1990.

[70] John Peter Wilson and John C Gallant. *Terrain analysis: principles and applications*. John Wiley & Sons, 2000.

[71] Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information science and systems*. John Wiley & Sons, 2015.

[72] Michael John De Smith, Michael F Goodchild, and Paul Longley. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Publishing Ltd, 2007.

[73] Jochen Schmidt, Ian S Evans, and Johannes Brinkmann. Comparison of polynomial models for land surface curvature calculation. *International Journal of Geographical Information Science*, 17(8):797–814, 2003.

[74] Helena Mitášová and Jaroslav Hofierka. Interpolation by regularized spline with tension: Ii. application to terrain modeling and surface geometry analysis. *Mathematical Geology*, 25(6):657–669, 1993.

[75] Joseph Wood. The geomorphological characterisation of digital elevation models. 1996.

[76] Ian Donald Moore, RB Grayson, and AR Ladson. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes*, 5(1):3–30, 1991.

[77] Ian S Evans. General geomorphometry, derivatives of altitude, and descriptive statistics. *Spatial analysis in geomorphology*, pages 17–90, 1972.

[78] Igor V Florinsky. Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, 12(1):47–62, 1998.

[79] Lyle W Zevenbergen and Colin R Thorne. Quantitative analysis of land surface topography. *Earth surface processes and landforms*, 12(1):47–56, 1987.

[80] D Chen*, DA Stow, and P Gong. Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. *International Journal of Remote Sensing*, 25(11):2177–2192, 2004.

[81] Markus Neteler and Helena Mitasova. *Open source GIS: a GRASS GIS approach*, volume 689. Springer Science & Business Media, 2013.

[82] E. W. Weisstein. Power sum. from mathworld–a wolfram web resource. `http://mathworld.wolfram.com/PowerSum.html`, April 2018.

[83] Eric W Weisstein. Power sum. 2002.

[84] International Water Institute. Red river basin decision information network. `https://iwinst.org/`, May 2017.

[85] Planet Imagery and Archive. Planet. `https://www.planet.com/products/planet-imagery/#re-imagery-product`, May 2017.

[86] Pierre Legendre. Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673, 1993.

[87] ESRI. Understanding a semivariogram: The range, sill, and nugget. `https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/understanding-a-semivariogram-the-range-sill-and-nugget.htm`, July 2019.

[88] Richard Webster and Margaret A Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, 2007.

[89] Harvey J Miller. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004.

[90] Peter M Atkinson and P Lewis. Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences*, 26(4):361–371, 2000.

[91] Jason W Karl. Spatial predictions of cover attributes of rangeland ecosystems using regression kriging and remote sensing. *Rangeland Ecology & Management*, 63(3):335–349, 2010.

[92] ESRI. Trend analysis. `https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/trend-analysis.htm`, July 2019.

[93] Carol A Gotway Crawford and Gary W Hergert. Incorporating spatial trends and anisotropy in geostatistical mapping of soil properties. *Soil Science Society of America Journal*, 61(1):298–309, 1997.

[94] Alex J Dumbrell, Ewen J Clark, Gillian A Frost, Thomas E Randell, Jonathan W Pitchford, and Jane K Hill. Changes in species diversity following habitat disturbance are dependent on spatial scale: theoretical and empirical evidence. *Journal of Applied Ecology*, 45(5):1531–1539, 2008.

[95] Alan H Strahler. The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote sensing of Environment*, 10(2):135–163, 1980.

[96] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[97] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[98] Rahul Gomes, Mostofa Ahsan, and Anne Denton. Random forest classifier in sdn framework for user-based indoor localization. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0537–0542. IEEE, 2018.

[99] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. ctree: Conditional inference trees. *The Comprehensive R Archive Network*, pages 1–34, 2015.

[100] Robert I Lerman and Shlomo Yitzhaki. A note on the calculation and interpretation of the gini index. *Economics Letters*, 15(3-4):363–368, 1984.

[101] Jeffrey S Evans, Melanie A Murphy, Zachary A Holden, and Samuel A Cushman. Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology*, pages 139–159. Springer, 2011.

[102] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[103] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. Chapman and hall/CRC, 2000.

[104] Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163, 1999.

[105] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[106] Allan Just, Margherita De Carli, Alexandra Shtein, Michael Dorman, Alexei Lyapustin, and Itai Kloog. Correcting measurement error in satellite aerosol optical depth with machine learning for modeling pm2. 5 in the northeastern usa. *Remote Sensing*, 10(5):803, 2018.

[107] R documentation. partialplot. `https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/partialPlot`, July 2018.

[108] AG RapidEye. Satellite imagery product specifications. *Satellite imagery product specifications: Version*, 2011.

[109] David M Mark. Network models in geomorphology. *Modelling Geomorphological Systems. John Wiley and Sons New York. 1988. p 73-97, 11 fig, 3 tab, 60 ref. NSF Grant SES-8420789.*, 1988.

[110] GIS Geography. What is ndvi (normalized difference vegetation index)? https://gisgeography.com/ndvi-normalized-difference-vegetation-index, July 2019.

[111] NASA Earth Observatory. Measuring vegetation. https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php, July 2019.

[112] Sentinel Hub by Sinergise. Ndvi (normalized difference vegetation index). https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index, July 2019.

[113] GIS Geography. How to create ndvi maps in arcgis. https://gisgeography.com/how-to-ndvi-maps-arcgis/, July 2019.

[114] Redlands ESRI. Arcgis desktop: release 10. *Environmental Systems Research Institute, CA*, 2011.

[115] R Core Team et al. R: A language and environment for statistical computing. 2013.

[116] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[117] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

[118] Dipti D Patil, VM Wadhai, and JA Gokhale. Evaluation of decision tree pruning algorithms for complexity and classification accuracy. *International Journal of Computer Applications*, 11(2):23–30, 2010.

[119] Nikita Patel and Saurabh Upadhyay. Study of various decision tree pruning methods with their empirical comparison in weka. *International journal of computer applications*, 60(12), 2012.

[120] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[121] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004.

[122] B Ghimire, J Rogan, and J Miller. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the getis statistic. *Remote Sensing Letters*, 1(1):45–54, 2010.

[123] Ugur Alganci, Baris Besol, and Elif Sertel. Accuracy assessment of different digital surface models. *ISPRS International Journal of Geo-Information*, 7(3):114, 2018.

[124] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[125] Papia Rozario, Buddhika Madurapperuma, and Yijun Wang. Remote sensing approach to detect burn severity risk zones in palo verde national park, costa rica. *Remote Sensing*, 10(9):1427, 2018.

[126] Papia F Rozario, Peter G Oduor, Larry Kotchman, and Michael Kangas. Uncertainty analysis of spatial autocorrelation of land-use and land-cover data within pipestem creek in north dakota. *Journal of Geoscience and Environment Protection*, 5(08):71, 2017.

[127] Nikhil R Garge, Georgiy Bobashev, and Barry Eggleston. Random forest methodology for model-based recursive partitioning: the mobforest package for r. *BMC bioinformatics*, 14(1):125, 2013.

[128] KK Datta and C De Jong. Adverse effect of waterlogging and soil salinity on crop and land productivity in northwest region of haryana, india. *Agricultural water management*, 57(3):223–238, 2002.

[129] NDSU Extension Service. Groundwater and its effect on crop production. `https://northcentralwater.org/files/2016/02/Groundwater-and-Its-Effect-on-Crop-Production.pdf`, July 2019.

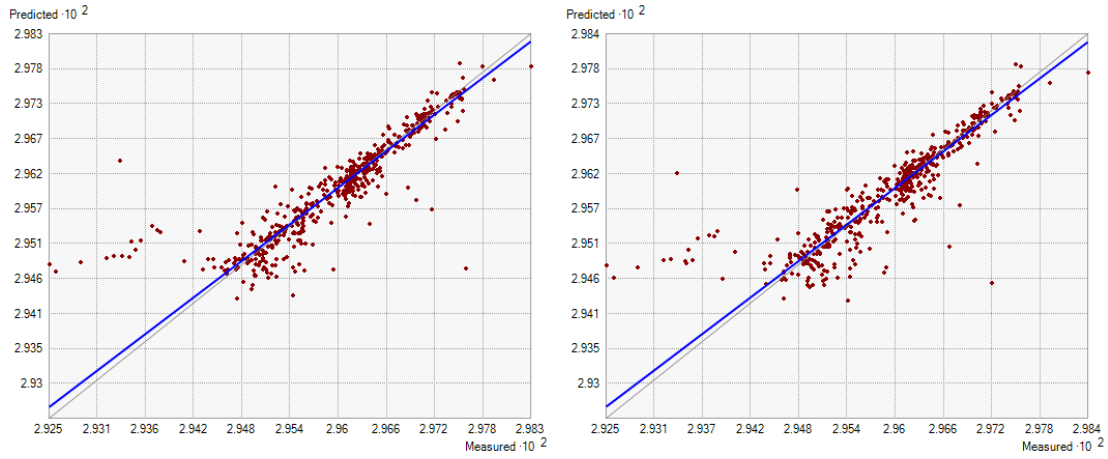
[130] Marco Vannucci and Valentina Colla. Meaningful discretization of continuous features for association rules mining by means of a som. In *ESANN*, pages 489–494. Citeseer, 2004.

[131] Richard Franke. Smooth interpolation of scattered data by local thin plate splines. *Computers & mathematics with applications*, 8(4):273–281, 1982.

[132] ESRI ArcGIS Pro. How spline works. `https://pro.arcgis.com/en/pro-app/tool-reference/3d-analyst/how-spline-works.htm`, July 2019.

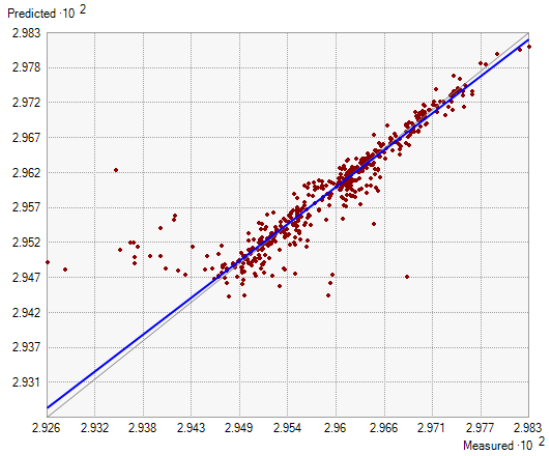
APPENDIX A. KRIGING CROSS-VALIDATION IN CHAPTER THREE

These plots summarize the observed vs the predicted values of DEMs that were derived after performing a semivariogram analysis for the study area in Chapter 3. A total of five hundred points were used to derive DEM simulations. These points used data from the DEMs obtained by multi-scalar analysis for $w = 4, 8, 16, 32$ and 64. Results for cross-validation were compared with values extracted from DEMs derived by ArcGIS processing for same set of five hundred points using window sizes $w = 3, 9, 15, 30$ and 63. The RMSE values obtained from the points in these plots have been summarized in Table 3.4. Both the approaches were able to generate models with comparable RMSE values. However, at larger window scales, especially $w = 64$ from proposed method (left) and $w = 63$ for ArcGIS processing (right) there was a comparable difference. For $w = 64$ in Appendix Figure A.1i the simulations predicted DEM values to be closer to their actual values. This was observed in the plot as the points lie close to the straight line compared to the plot in Appendix Figure A.1j for $w = 63$ where the points are scattered far apart showing more deviation between observed and predicted values.

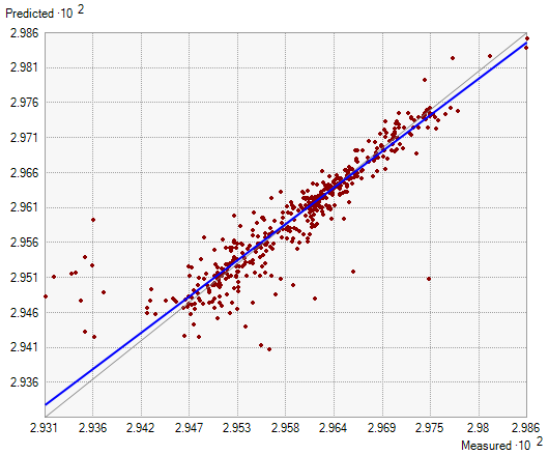


(a) $w = 4$ (b) $w = 3$

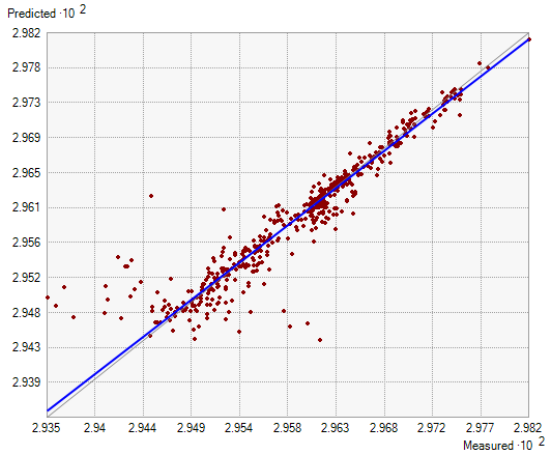
Figure A.1. Kriging cross validation results for proposed aggregation (left) and ArcGIS approach (right)



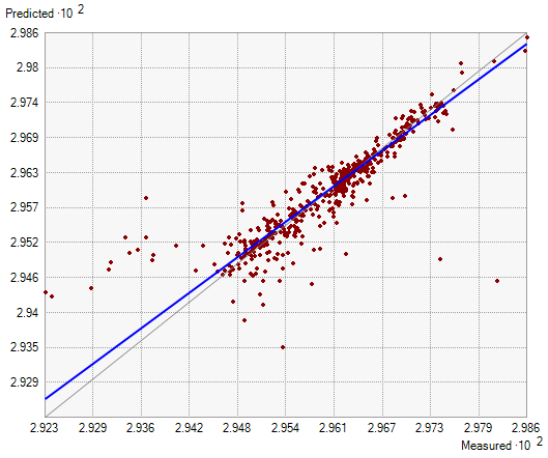
(c) $w = 8$



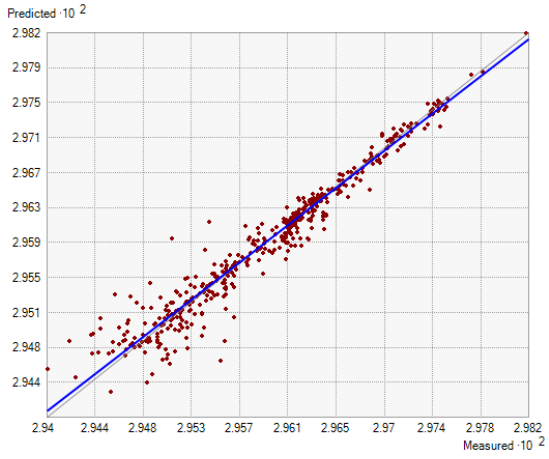
(d) $w = 9$



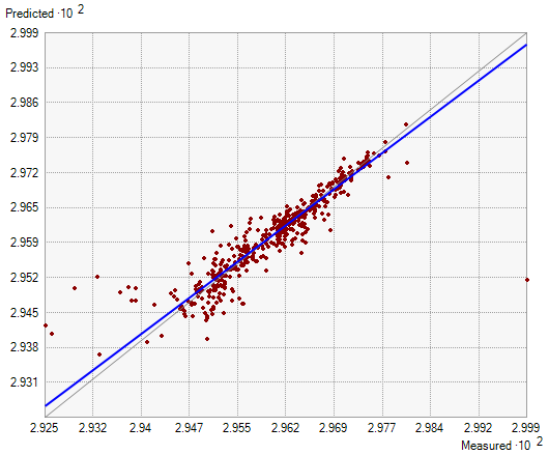
(e) $w = 16$



(f) $w = 15$

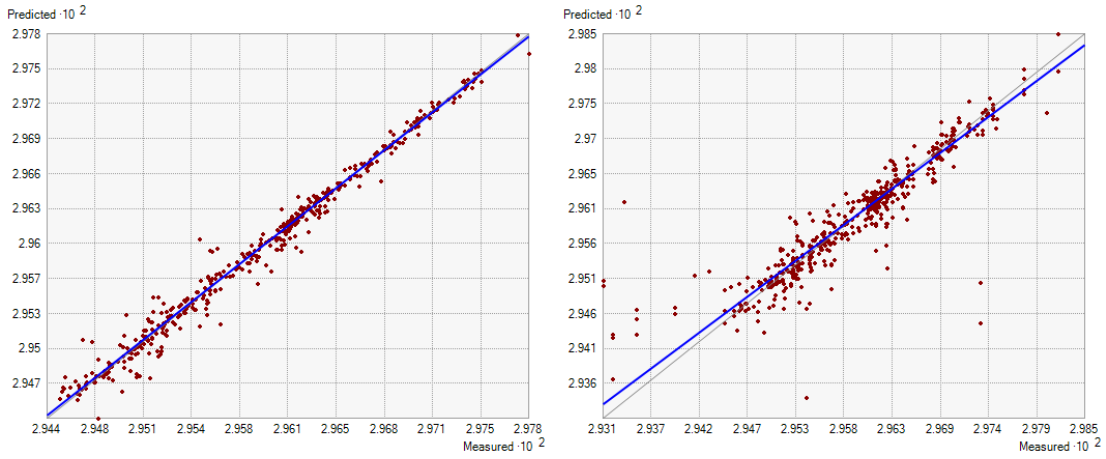


(g) $w = 32$



(h) $w = 30$

Figure A.1. Kriging cross-validation results for proposed aggregation (left) and ArcGIS approach (right)(continued)



(i) $w = 64$ (j) $w = 63$

Figure A.1. Kriging cross-validation results for proposed aggregation (left) and ArcGIS approach (right)(continued)

APPENDIX B. R SCRIPTS USED IN CHAPTER FOUR

The following R script shows how the Random Forest and Naive Bayes classification model were generated in Chapter 4 of the dissertation. The results from the DEM, curvature, slope, aspect and NDVI were exported to R in lines 9, 15, and 21. These text files contained the elevation, slope, aspect, curvature and NDVI values for the same study area. The five different text files contained records corresponding to same study area in different resolutions $w = 4, 8$, and 16.

```
1  #Import libraries
2  library(randomForest)
3  library(pdp)
4  library(plyr)
5  library(party)
6  library(naivebayes)
7  library(caret)
8  # All the A's contain a copy of the original file.
9  A_4<-read.table('Points/W4.txt',header=TRUE,sep=',')
10 # Remove excess columns.
11 A_4 $grid_code<-NULL
12 A_4 $pointid<-NULL
13 A_4 $FID<-NULL
14 A_4 $PrCurv_W4<-NULL
15 A_4 $PlCurv_W4<-NULL
16 # Read from a text file.
17 A_8<-read.table('Points/W8.txt',header=TRUE,sep=',')
18 A_8$grid_code<-NULL
19 A_8$pointid<-NULL
20 A_8$FID<-NULL
21 A_8$PrCurv_W8<-NULL
```

```
A_8$PlCurv_W8<-NULL
A_16<-read.table('Points/W16.txt',header=TRUE,sep=',')
A_16$grid_code<-NULL
A_16$pointid<-NULL
A_16$FID<-NULL
A_16$PrCurv_W16<-NULL
A_16$PlCurv_W16<-NULL
# All the B's Classify the dataset using positive curvature.
B_4<-subset(A_4,A_4$Curv_W4>=0)
B_4$NDVI_W4<-round(B_4$NDVI_W4,4)
B_8<-subset(A_8,A_8$Curv_W8>=0)
B_8$NDVI_W8<-round(B_8$NDVI_W8,4)
B_16<-subset(A_16,A_16$Curv_W16>=0)
B_16$NDVI_W16<-round(B_16$NDVI_W16,4)
# All the C's aggregate the results according to average NDVI
C_4<-aggregate(. NDVI_W4,data=B_4,mean)
C_8<-aggregate(. NDVI_W8,data=B_8,mean)
C_16<-aggregate(. NDVI_W16,data=B_16,mean)
#Copy to a new set
C1_4<-C_4
C1_8<-C_8
C1_16<-C_16
str(C1_4)
# Grouping the dataset based on classes
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.3 & C1_4$NDVI_W4 < 0.4] <- '1'
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.4 & C1_4$NDVI_W4 < 0.5] <- '2'
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.5 & C1_4$NDVI_W4 < 0.6] <- '3'
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.6 & C1_4$NDVI_W4 < 0.7] <- '4'
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.7 & C1_4$NDVI_W4 < 0.8] <- '5'
```

```
C1_4$NDVI_W4[C1_4$NDVI_W4 >= 0.8 & C1_4$NDVI_W4 < 0.9] <- '6'
C1_4$NDVI_W4<-as.factor(C1_4$NDVI_W4)
# Partition to training and testing
set.seed(1234)
pd_4 <-sample(2,nrow(C1_4), replace=TRUE, prob = c(0.6,0.4))
train_4 <- C1_4[pd_4 ==1,]
validate_4 <- C1_4[pd_4 ==2,]
#Random Forest Classification and prediction
Rf_4 <- randomForest(NDVI_W4 ., data=train_4, ntree = 500, importance = TRUE,proximity
= TRUE)
Rf_4
importance(Rf_4)
Rf_4_Predict <- predict(Rf_4, validate_4)
confusionMatrix(Rf_4_Predict, validate_4$NDVI_W4)
#Copy to a new set
C1_4_Rg<-C_4
C1_8_Rg<-C_8
C1_16_Rg<-C_16
str(C1_4_Rg)
# Naive Bayes Model and prediction
NB_4 <- naive_bayes(NDVI_W4   ., data = train_4, usekernel = T)
NB_4
NB_4_Predict <- predict(NB_4, validate_4)
confusionMatrix(NB_4_Predict, validate_4$NDVI_W4)
 Repeat the process for w = 8 and 16
```

This entire script was run once again for results derived from ArcGIS with window scales $w = 3, 9,$ and 15 that were exported in different text files.

APPENDIX C. R SCRIPTS USED IN CHAPTER FIVE

The following R script shows how the Random Forest based regression model and partial dependence plots were generated in Chapter 5 of the dissertation. The results from the DEM, curvature, slope, aspect and NDVI were exported to R in lines 6, 10, 14, 18, and 22. These text files contained the elevation, slope, aspect, curvature and NDVI values for the same study area. The five different text files contained records corresponding to same study area in different resolutions $w = 4, 8, 16, 32$, and 64.

```
1  #import libraries
2  library(randomForest)
3  library(pdp)
4  library(plyr)
5  # All the A's contain a copy of the original file.
6  A_4<-read.table('StudyArea/Points/W4.txt',header=TRUE,sep=',')
7  # Remove excess columns.
8  A_4$grid_code<-NULL
9  A_4$pointid<-NULL
10 A_4$FID<-NULL
11 # Read from a text file.
12 A_8<-read.table('StudyArea/Points/W8.txt',header=TRUE,sep=',')
13 A_8$grid_code<-NULL
14 A_8$pointid<-NULL
15 A_8$FID<-NULL
16 A_16<-read.table('StudyArea/Points/W16.txt',header=TRUE,sep=',')
17 A_16$grid_code<-NULL
18 A_16$pointid<-NULL
19 A_16$FID<-NULL
20 A_32<-read.table('StudyArea/Points/W32.txt',header=TRUE,sep=',')
```

```
A_32$grid_code<-NULL
A_32$pointid<-NULL
A_32$FID<-NULL
A_64<-read.table('StudyArea/Points/W64.txt',header=TRUE,sep=',')
A_64$grid_code<-NULL
A_64$pointid<-NULL
A_64$FID<-NULL
# All the B's Classify the dataset using positive curvature.
B_4<-subset(A_4,A_4$Curv_W4>=0)
B_4$NDVI_W4<-round(B_4$NDVI_W4,4)
B_8<-subset(A_8,A_8$Curv_W8>=0)
B_8$NDVI_W8<-round(B_8$NDVI_W8,4)
B_16<-subset(A_16,A_16$Curv_W16>=0)
B_16$NDVI_W16<-round(B_16$NDVI_W16,4)
B_32<-subset(A_32,A_32$Curv_W32>=0)
B_32$NDVI_W32<-round(B_32$NDVI_W32,4)
B_64<-subset(A_64,A_64$Curv_W64>=0)
B_64$NDVI_W64<-round(B_64$NDVI_W64,4)
# All the C's aggregate the results according to average NDVI
C_4<-aggregate(. NDVI_W4,data=B_4,mean)
C_8<-aggregate(. NDVI_W8,data=B_8,mean)
C_16<-aggregate(. NDVI_W16,data=B_16,mean)
C_32<-aggregate(. NDVI_W32,data=B_32,mean)
C_64<-aggregate(. NDVI_W64,data=B_64,mean)
# All the D's have random forest
D4_rf<-randomForest(NDVI_W4 .,data=C_4,importance=TRUE)
D8_rf<-randomForest(NDVI_W8 .,data=C_8,importance=TRUE)
D16_rf<-randomForest(NDVI_W16 .,data=C_16,importance=TRUE)
D32_rf<-randomForest(NDVI_W32 .,data=C_32,importance=TRUE)
```

```
D64_rf<-randomForest(NDVI_W64 .,data=C_64,importance=TRUE)
#All the E's contain the partial dependence plots for w = 4,8, 16, 32 and 64
E1_4<-partial(D4_rf, pred.var = "Curv_W4", plot = TRUE, rug = TRUE)
E2_4<-partial(D4_rf, pred.var = "Slope_W4", plot = TRUE, rug = TRUE)
E3_4<-partial(D4_rf, pred.var = c("Curv_W4", "Slope_W4"), plot = TRUE, chull = TRUE)
pd <- partial(D4_rf, pred.var = c("Curv_W4", "Slope_W4"))
E3_5 <-plotPartial(pd, levelplot = FALSE, zlab = "NDVI_W4", drape = TRUE, colorkey =
TRUE, screen = list(z = -120, x = -60)) E1_8<-partial(D8_rf, pred.var = "Curv_W8", plot =
TRUE, rug = TRUE)
E2_8<-partial(D8_rf, pred.var = "Slope_W8", plot = TRUE, rug = TRUE)
E3_8<-partial(D8_rf, pred.var = c("Curv_W8", "Slope_W8"), plot = TRUE, chull = TRUE)
E1_16<-partial(D16_rf, pred.var = "Curv_W16", plot = TRUE, rug = TRUE)
E2_16<-partial(D16_rf, pred.var = "Slope_W16", plot = TRUE, rug = TRUE)
E3_16<-partial(D16_rf, pred.var = c("Curv_W16", "Slope_W16"), plot = TRUE, chull =
TRUE)
E5_16<-partial(D16_rf, pred.var = c("Elev_W16","Curv_W16"), plot = TRUE, rug=TRUE)
E1_32<-partial(D32_rf, pred.var = "Curv_W32", plot = TRUE, rug = TRUE)
E2_32<-partial(D32_rf, pred.var = "Slope_W32", plot = TRUE, rug = TRUE)
E3_32<-partial(D32_rf, pred.var = c("Curv_W32", "Slope_W32"), plot = TRUE, chull =
TRUE)
E5_32<-partial(D32_rf, pred.var = c("Elev_W32","Curv_W32"), plot = TRUE, rug = TRUE)
E1_64<-partial(D64_rf, pred.var = "Curv_W64", plot = TRUE, rug = TRUE)
E2_64<-partial(D64_rf, pred.var = "Slope_W64", plot = TRUE, rug = TRUE)
E3_64<-partial(D64_rf, pred.var = c("Curv_W64", "Slope_W64"), plot = TRUE, chull =
TRUE)
options(scipen=10000)
# Do RMSE for ElevValues. All the F's have Elevation
 F_1<-read.table('ElevValues.txt',header=TRUE,sep=',')
F_1$FID<-NULL
```

```
F_1$CID<-NULL
F_1$Error4<-F_1$W0-F_1$W4
F_1$Error8<-F_1$W0-F_1$W8
F_1$Error16<-F_1$W0-F_1$W16
F_1$Error32<-F_1$W0-F_1$W32
F_1$Error64<-F_1$W0-F_1$W64
F_1$Error4sq<-F_1$Error4*F_1$Error4
F_1$Error8sq<-F_1$Error8*F_1$Error8
F_1$Error16sq<-F_1$Error16*F_1$Error16
F_1$Error32sq<-F_1$Error32*F_1$Error32
F_1$Error64sq<-F_1$Error64*F_1$Error64
RMSE4<-sqrt(sum(F_1$Error4sq)/800)
RMSE8<-sqrt(sum(F_1$Error8sq)/800)
RMSE16<-sqrt(sum(F_1$Error16sq)/800)
RMSE32<-sqrt(sum(F_1$Error32sq)/800)
RMSE64<-sqrt(sum(F_1$Error64sq)/800)
# Repeat the process for negative curvature values by replacing the >= symbol with =< sym-
bol in lines 27 to 35

```

This entire script was run once again for results derived from ArcGIS with window scales $w = 3, 9, 15, 30$, and 63 that were exported in different text files.

APPENDIX D. PUBLISHED WORK RELATED TO THIS RESEARCH

Anne, Denton, Rahul Gomes, and David Franzen. "Scaling up Window-Based Slope Computations for Geographic Information System." 2018 IEEE International Conference on Electro/Information Technology (EIT). IEEE, 2018. [49]

Rahul, Gomes, Anne Denton, and David Franzen. "Quantifying Efficiency of Sliding-Window Based Aggregation Technique by Using Predictive Modeling on Landform Attributes Derived from DEM and NDVI." ISPRS International Journal of Geo-Information 8.4 (2019): 196. [50]

Rahul, Gomes, Anne Denton and David Franzen. "Comparing classification accuracy of NDVI with DEM derived attributes using multi-scalar approach in Geographic Information Systems." 2019 IEEE International Conference on Electro/Information Technology (EIT). IEEE, 2019 [48]

Anne, Denton, Rahul Gomes, and David Franzen. "Large-window Curvature Computations for High-resolution Big Geospatial Data." In progress

Denton Anne, Gomes Rahul, and Franzen David. "Separating Landform from Noise in High-Resolution Digital Elevation Models through Scale-Adaptive Window-Based Regression" International Conference on Spatial Statistics and Geostatistics (ICSSG 2019), New York, NY

Gomes Rahul, and Denton Anne, "Incorporating Data Mining and Iterative Aggregation on Geospatial Datasets to Understand Soil Health in Depressions." ND EPSCoR 2019 State Conference, Fargodome, Fargo, ND

Gomes Rahul, and Denton Anne. "Taking Terrain Analysis to the Big Data Era for Understanding Soil Health in Depressions." ND EPSCoR 2018 State Conference, Alerus center, Grand Forks, ND