LANDSLIDE FORECASTING: INVENTORY, SUSCEPTIBILITY, AND HAZARD ANALYSES

By

Yashar Alimohammadlou

A Dissertation
Submitted to the
Graduate Faculty
Of
George Mason University
In Partial Fulfillment of
The Requirements for Degree
Of
Doctor of Philosophy
Civil and Infrastructure Engineering

Committee:

______________________________ Dr. Burak F. Tanyu, Dissertation Director

______________________________ Dr. Girum Urgessa, Committee Member

______________________________ Dr. Viviana Maggioni, Committee Member

______________________________ Dr. Paul V. Lade, Committee Member

______________________________ Dr. Gheorghe Tecuci, Committee Member

______________________________ Dr. Sam Salem, Department Chair

______________________________ Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: ________________________

Spring Semester 2020
George Mason University
Fairfax, VA

Landslide Forecasting: Inventory, Susceptibility, and Hazard Analyses

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Yashar Alimohammadlou
Master of Science
Azad University of Zanjan, 2012
Bachelor of Science
Azad University of Tabriz, 2009

Director: Burak F. Tanyu, Associate Professor
Department of Civil, Environmental and Infrastructure Engineering

Spring Semester 2020
George Mason University
Fairfax, VA

DEDICATION

To Those Who Seek the Truth

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Burak Tanyu for his excellent academic guidance, material support, and technical and social discussions in my research.

TABLE OF CONTENTS

Contents Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Automated Landslide Detection Model	ALDM
Digital Elevation Model	DEM
Feature Selection Method	FSM
Geographical Information Systems	GIS
Green-Ampt	GA
Landslide Susceptibility Analysis	LSA
Light Detection and Ranging	LiDAR
Pennsylvania Department of Conservation and Natural Resources	PADCNR
Pennsylvania Spatial Data Access	PASDA
Stability Index Mapping	SINMAP
Shallow Landslide Stability Model	SHALSTAB
Transient Rainfall Infiltration and Grid based Regional Slope stability	TRIGRS
United States Geological Survey	USGS
United States Department of Agriculture	USDA
Visual Detection	VD

ABSTRACT

LANDSLIDE FORECASTING: INVENTORY, SUSCEPTIBILITY, AND HAZARD ANALYSES

Yashar Alimohammadlou, Ph.D.

George Mason University, 2020

Dissertation Director: Dr. Burak F. Tanyu

The effects of landslides have been exponentially increasing due to the rapid growth of urbanization and global climate change. However, the methods to evaluate the effects of landslides on very large areas (such as an entire County or State) are still very limited. Forecasting landslides is a complex process, which involves (at a minimum) three different steps. First, inventory maps of the area that documents the extent of the existing landslides must be developed. Second, analyses must be conducted to understand the common features within the areas of landslides. Third, the susceptibility of the areas that may or may not have landslides must be determined at the moment to have the potential to be part of a landslide in the future (identifying the zones of high potential areas). Once the inventory maps and susceptibility analyses are completed, then hazard forecasting analyses must be conducted to evaluate the effects of external (triggering) factors on initiating future

landslides both within the areas that had past landslides and no landslides. The research presented in this dissertation focused on studying all of these processes (steps) and provide new methods. The outcome of the research was to create a complete methodology of landslide forecasting that may be used by decision makers (such as Counties, State agencies, insurance companies, etc.) to evaluate large regions and use this information to help communities in terms of the potential dangers that may exist now and in the future as it relates to landslides. The results of this study are utilized to develop a complete and reliable framework of inventory, susceptibility, and hazard analyses. Consequently, the outcome of this research may serve as an early warning system for landslide prone and at-risk regions throughout the world.

CHAPTER 1: RESEARCH OVERVIEW

Landslides are mass movements that may consist of soil, rock, debris, organic matter, artificial fill, or combination of these materials and may be triggered by a variety of factors, such as rainfall (precipitation), seismic activity, ground water level change, storm waves or rapid stream erosion. These triggering factors cause a rapid increase in shear stress or decrease in shear strength of slope-forming materials. Combination of several different sources of data indicate that worldwide, landslides cause losses of around 1,150 lives and $4 billion economic damage annually (IDD, 2018; WHO, 2017; Pardeshi et al., 2013). In the United States alone, annually, the economic losses from landslides were estimated as $1–2 billion and loss of lives as 50 (Based on the data from USGS). The number of loss of lives in the U.S. caused by landslides were estimated to be more than the average deaths caused by earthquakes (Highland et al., 2008). Additionally, Dilley et al. (2005) show that more than 300 million of the world’s population is exposed to landslides and 66 million people are living in high risk areas. Predicting landslides and analyzing potential risks can have major importance in minimizing losses from these natural hazards. Studying landslide risk will determine sufficient measures in different regions for improvement in hazard recognition, prediction, mitigation measures, warning systems, hazard mapping, and emergency preparedness response and recovery.

Terzaghi (1950) illuminated that landslide occurrence is dependent on the distinction between internal changes within a slope; for example, factors that induce shear strength reduction, and external causes, which give rise to an increased shear stress. It is necessary to note that the subaerial mass movements are mostly triggered by precipitation (rainfall and snowmelt), seismic activity, human interventions, volcanism, weathering, seepages and springs, and river erosion. There is no single factor that can accurately characterize landslide hazard; a collection of parameters should be analyzed interactively. Based on this in 1989, Crozier developed a classification and arranged influential parameters into four categories as (1) precondition (predisposing) factors, (2) conditioning factors, (3) triggering factors, and (4) sustaining factors.

The first group includes inherent slope parameters which may not have direct effects but can precipitate the slide process as a catalyst (for example, soil or rock material characteristics). In the second group, preparatory factors are dynamic parameters, which reduce the margin of stability over time without initiating movement and as noted before, would be accelerated by precondition factors. The third group of this classification constitutes triggering factors, which initiate slope movement and initiate sliding. Most of the triggering parameters are external forces, and the initiating thresholds are thus referred to as extrinsic thresholds (Schumm, 1979). The sustaining factors cover the parameters which predicate the slides' behavior such as rate, duration, and form of movement.

Landslide studies require a minimum of three steps:

- ***Inventory maps*** which classify and map the documentation of existing landslides. They can be either recently occurring or historic and pre-historic events (Barlow et al., 2003; Glenn et al., 2006; Leshchinsky et al., 2015; Sato et al., 2007);
- ***Susceptibility analysis*** which starts with landslide inventory maps and evaluates the soil and site conditions that indicate areas with high landslide potential. The results of this analysis can be projected into a landslide susceptibility map (Chen et al., 2017; Pham et al., 2017; Yeon et al., 2010);
- ***Hazard forecasting*** which can be prepared by incorporating the external (triggering) factors (Baum & Godt, 2010; Dhakal & Sidle, 2004; Rossi et al., 2013). The most important difference between the hazard forecasting and susceptibility analysis is that the triggering sources are also included in the hazard forecasting.

Investigation of landslide risk will determine sufficient measures in different regions for improvement in hazard recognition, prediction, mitigation measures, warning systems, hazard mapping and assessments, and emergency preparedness response and recovery. Development of landslide studies based on attaining a reliable framework of inventory, susceptibility, hazard analyses, and consequently an early warning system will be useful for landslide prone regions and at-risk countries. Based on landslide studies, many countries can be able to adopt powerful economic policies in order to support mitigation strategies against this hazard. On the other hand, development of landslide studies in countries with high frequency of occurrence and attaining a reliable framework will be useful for other regions. Consequently, developing countries will be able to adopt powerful economic policies in order to support mitigation strategies against this hazard.

Introduction to Problem Statement

In general, there are two main groups of researchers who deal with landslide studies: geotechnical engineers and geologist/geography researchers. Each group utilizes a different approach in studying different aspects of landslide studies. Slope stability analysis is a method used by geotechnical engineers to estimate the potential failure mechanism of human-made or natural slopes (e.g., excavations, landfills, road-cuts, and landslides). The calculation of slope stability requires precise data in failure surface (slope cross sections) from field observations and topographical maps, soil shear strength parameters, and pore water pressure from multiple in-situ geotechnical tests. Therefore, one of the main disadvantages of this approach is that slope stability is evaluated by determining a worst-case scenario where slope and the associated parameters and features are drawn as part of a specific cross section in great detail. Additionally, the factor of safety is computed individually for each cross section of the slope. Thus, predicting shallow landslides over large regions in real or near real time is an impossible task based on the deterministic approach (Baum & Godt, 2010). In other words, geotechnical engineers are able to conduct specific evaluations of a given slope. Therefore, the present study made an effort to expand this capability at a regional level (covering very large areas such as an entire County within a region).

The second group is geological and geographical scientists who look at the problem in a big and regional scale. While geotechnical engineers are more interested in the index and engineering properties of soils and their relationship with the failure scenarios, geologists and geographers are more concentrated on topographical and geomorphological properties such as slope aspect, curvature, wetness index, distance to drainage, distance to fault, drainage density, fault density, and spring density maps (Ozdemir & Altural, 2013; Hong et al., 2015; Trigila et al., 2015; Youssef et al., 2016; Wu et al., 2020). On the other hand, the conventional approach for landslide inventory analyses involves using the data obtained through the field survey mapping and historical records, which is tedious to compile especially in heavily vegetated terrains (Pardeshi et al., 2013). New advancements in satellite imagery, LiDAR data acquisition, and GIS systems improve the landslide inventory techniques in different ways. However, all previously developed models utilize a supervised method to delineate existing landslides in a selected region. This means that the supervised method requires the user/operator to predefine landslide geometrical features (scarp, toe, and height of body) in a selected region. Moreover, these studies demonstrate the ability to detect existing landslides with well-defined features. In regions such as east coast of the U.S., landslides are often smaller in scale and sporadic; further, their features such as scarp and toe tend to be ill-defined. In these conditions, the previously determined landslide detection algorithms are not as effective because they require supervision wherein the user is expected to define features of the landslides (e.g., the height of the scarp, the slope of the body, and the toe of the landslide) in a given region.

The purpose of this study was to develop a framework of inventory, susceptibility, hazard analyses, using Digital Elevation Model (DEM) data and landslide conditioning and triggering factors obtained from available resources. The study proposes a methodology to evaluate the risk of landslide prone areas in a given region by performing inventory, susceptibility, and hazard analyses. Using this framework, the results for existing landslides and inventory analyses will be provided by quickly running the model on DEM data obtained for the selected study area.

Outlining Gaps in the Literature - Brief Summary

Landslide inventory maps are the simplest form of landslide information, which records the location, the date of occurrence, and type of landslides (Cruden, 1991). The results of the study by Galli et al. (2008) revealed that complete landslide inventory map provide high predictive power for further landslide susceptibility and hazard analyses. The conventional techniques such as field survey mapping and historical records are difficult to make for various reasons such as the size of the landslide, the viewpoint of the investigator, and the fact that old landslides are often partially or totally covered by forest, or have been partly dismantled by other landslides. One of the previously used methods in landslide inventory analysis is the interpretation of airborne or satellite imagery (Mwaniki et al., 2017). The most important challenge with using satellite imagery is to interpret the observed features in areas with heavy vegetation cover. The main technique used in interpreting satellite imagery involves in recognizing the difference in reflectance and shape of multiple colors and this can cause significant errors since many existing landslides are covered by dense vegetation. The use of light detection and ranging (LiDAR) data to create bare earth digital elevation models (DEM) is an effective means of mapping existing landslides in a given region (Barlow et al., 2003; McKean and Roering, 2004; Glenn et al., 2006; Booth et al., 2009; Leshchinsky et al. 2015; Li et al., 2016). One of the challenges with previously developed inventory analyses is that they require experience, training, a systematic methodology, and well-defined interpretation that relies on the expert's opinion (Guzzetti et al., 2012). In regions such as east coast of the U.S. this becomes an issue since landslides are often smaller in scale and sporadic; further, their features such as scarp and toe tend to be not as well-defined.

Landslide susceptibility analysis is the second step of the whole forecasting approach and is based on evaluating the likelihood of the occurrence based on the prelimina ry (or preparatory) factors. There are three components that determine the result of landslide susceptibility analyses: 1) features and dataset combination, 2) landslide observations, and 3) algorithms used for evaluation. Change in these components may affect the accuracy of landslide susceptibility analyses or the ability of landslide forecasting. In general, the parameters used for landslide susceptibility analysis can be divided into categories, such as, geological, hydrological, topographical, and geotechnical features. However, all of the previous studies have mostly used geological and topographical parameters of the study area but have not included the geotechnical and hydrological parameters. In previous work, the observations are generally modeled as vector or point data on the area (Pham et al., 2017b). This means that in LSA model, each landslide occurrence is represented as a point on a map and in the dataset. The modeling of the landslide observations also captures the proportions of the landslides within an area. Most of the studies have used this data to select areas for their analyses where proportion of landslide and non-landslide areas is equal to each other (Ozdemir and Altural, 2013, Chen et al., 2017). The primary reasons for this approach are due to the ease of the computational efforts and the sensitivity of the assessment methods to the unbalance datasets. Previous studies have used multiple algorithms to perform susceptibility analyses. Frequency Ratio (Hong et al., 2015; Ozdemir and Altural, 2013), Logistic Regression (Oh et al., 2010; Ozdemir and Altural, 2013; Youssef et al., 2016), Random Forest (Chen et al., 2017; Trigila et al., 2015), Classification and Regression Tree (Chen et al., 2017), and Support Vector Machines (Colkesen et al., 2016) are the methods that have been previously used for this purpose. However, due to the fact that using raster dataset would require algorithms that are not sensitive to unbalanced data, the present study utilized C4.5 and C5.0 algorithm to evaluate the landslide susceptibility analyses. The results of this analyses were compared against Random Forest analyses which is a standard method used previously by multiple studies.

Hazard analyses consist of mapping and evaluating the potential for damage by incorporating external factors. Several hazard analyses approaches have been utilized previously to address landslide hazard analysis. SHALSTAB (Shallow Landslide Stability Model) is one of the grid-based approaches. However, this model is not appropriate to forecast the timing of slope failure based on the triggering mechanism (Dhakal and Sidle, 2004). SINMAP (Stability Index Mapping) is a different method that works based on infinite slope stability model with groundwater pore pressures (Rossi et al., 2013). This method applies only to shallow transitional landslide phenomena controlled by shallow groundwater flow. It does not apply to deep-seated or rotational slides. Another disadvantage of this method is that the accuracy of output is heavily reliant on the accuracy of DEM (Yatheendradas et al., 2019). TRIGRS (Transient Rainfall Infiltration and Grid based Regional Slope stability) is another approach developed in FORTRAN language, for computing the transient pore pressure distribution due to rainfall infiltration using the method proposed by (Iverson, 2000) (Baum and Godt, 2010. However, TRIGRS is very sensitive to initial conditions, therefore, if the initial water table depth is poorly constrained, it may produce questionable results. All of the above discussed approaches have a value to estimate landslide hazard analyses but they do not consider geotechnical/geological features within the landslide area (e.g., types of bedrock and soil, depths of stratigraphic layers, percentage of soils, and shear strength parameters of soil). Therefore, the analyses are predominantly solved by using rainfall data and topographical information of the slope alone, ignoring the rest of the very important properties. This approach potentially grossly limits the accuracy of the predictions.

In summary, the following challenges remain unaddressed:

- Detecting landslides at a regional scale without predefining geometrical features (e.g., the location of the scarp, toe, etc.) and using a technique that does not require supervision (e.g. revising the resolution of the scarp, or defining a minimum slope for the body of the landslide, etc.) to perform the analyses.
- When conducting landslide prediction and hazard analyses, the ability to incorporate not only the geometrical features but also the geotechnical parameters to define the properties of the region.
- Datasets created in previous landslide analyses are based on compiling data based on vector, which results in a point within a region for each location of interest. The analyses are conducted by creating equal amount of vector data to represent

landslide and non-landslide locations. However, in reality, in a given region, the areas covered by landslides are typically not equal to the areas covered by non-landslides (i.e., referred as unbalanced datasets). This is best modeled with creating raster type data sets as opposed to vector type. The ability to model a region with raster dataset requires rigorous analyses and such path is not explicitly discussed in the previous literature.

- For a given region that is well studied, predicting the future occurrence of a landslide based on a triggering factor (i.e., rainfall event).

The goal of this research was to fill in the gap for the above-mentioned challenges and achieve the following objectives:

1. Being able to develop a bare earth digital elevation model based on available LiDAR data and implement image segmentation method that is embedded in the existing and available GIS platform to very quickly and accurately map out the location of previously occurred landslides. This tool is envisioned to be as part of any initial site investigation efforts even in regions where the aerial photography cannot be used due to heavy vegetation cover.
2. Develop a methodology to be able to create raster datasets that allow the region to be evaluated with millions of pixels (as opposed to vector data with single pixel representing given condition (being landslide or non-landslide)) where each pixel contains multiple dataset including parameters associated with both geotechnical and geological properties. The applicability of the methodology to be evaluated with select existing algorithms where the results from unbalanced and balanced

datasets are compared. The end result of this objective is envisioned to outline whether the end users should go through the extensive effort to evaluate the regions with raster datasets or is the existing way of modeling with vector datasets are considered sufficient enough. Such decision will either validate the approaches followed by previous studies or will demonstrate the significant shortcomings of the previous studies.

3. Develop a methodology to estimate the probability of failure of previously existing landslides in a region based on infiltration of rainwater into the ground. Such tool is envisioned to be used by decision makers to develop early warning systems to identify potential conditions and areas that could be a treat for loss of property and life. The tool is not envisioned to be used for design but can also be used as part of site investigation to characterize the land prior to any decisions for development.

Research Questions

Based on the extensive survey of present literature in inventory, susceptibility, and hazard analyses and by using the initial results of landslide forecasting models, four different research questions were developed in order to evaluate and test the validity of the proposed hypotheses. These research questions are as follows:

Hypothesis 1: Can we detect distributed landslides using LiDAR (DEM) and image segmentation method without considering predefined geometrical parameters (scarp and toe)?

Hypothesis 2: Is it possible and more applicable to develop a methodology to combine all parameters (Geotechnical and Geomorphological) in one dataset when evaluating landslide susceptibility?

Hypothesis 3: If the landslide susceptibility analyses were performed by creating a raster (unbalanced) dataset, would the analyses be more accurate especially if such analyses were conducted with algorithms that are less sensitivity to dataset balance (such as C4.5 (J48) and C5.0?

Hypothesis 4: Can we estimate the probability of future landslides in large areas if the modeled developed for susceptibility analyses were modified to evaluate the triggering of landslides based on rainfall infiltration?

Overall Summary of the Conducted Research

In line with the motivation of the research, this study was conducted in three parts. Overall summary of each part is described below, following with sections of the dissertation that were written to present the details of each part.

In the first part of the research, a new tool that is herein referred as Automated Landslide Detection Model (ALDM) has been developed. This tool is then tested within three different areas in Pennsylvania (one of the most landslide prone regions in the east coast of the United States) where the state has an extensive inventory of landslides created by the Pennsylvania Department of Conservation and Natural Resources (PDCNR) in 2001. This dataset was created based on the traditional methods that involves surveying the area in the field (in-person) and interpretation of the available photogrammetric data. This approach works very well but it requires great amount of time, personnel, and budget as covering a large area in detail requires may people to be involved for many weeks and months. Also, although the observations at the time become absolute, any new landslide area that develops after these efforts are not incorporated into this database. The ALDM analyses allow such evaluation to be completed in a matter of days. To test the validity of the developed ALDM in this study, LiDAR bare earth digital elevation models (DEMs) of three different areas within PDCNR inventory have been created and the ALDM analyses were performed on this dataset. Additionally, using the DEM of the area, Hill shade map of these three study areas were also created and following a similar approach as conducted for the photogrammetric data evaluation, landslides were visually identified. Therefore, three datasets from three different areas within PA could be compared against each other

(PA DCNR field data, visual interpretation of Hill shape map, and ALDM predictions). The results demonstrated that the new ALDM method was able to accurately capture both the landslides and non-landslides in all of the areas evaluated with accuracies of 70% and 92% respectively when the result is compared against PA DCNR field data. The rate of accuracy is even higher when it is compared to visual detection method (86% for landslide and 94% for non-landslide areas). Additionally, the study showed that the proposed ALDM method could be implemented in different regions where landslides of different shapes, sizes, and abundance could be detected. The significance of the newly developed ALDM becomes evident when compared to the other existing computer-based analyses. The previously developed computer-based analyses require the user to obtain the specific code, to define the properties of the landslide prior to the search and require supervision and quality assurance. The newly developed ALDM can be implemented with readily available tools and datasets and does not need supervision for the analyses after the parameters regarding landslide morphology are defined for that region. Findings obtained from this part of the study allowed the confirmation of the Hypothesis 1. Details of this part of the study have been presented in Chapter 2.

In the second part of the research, the focus was to conduct Landslide Susceptibility Analysis (LSA) that could be used to develop warning systems and mitigation measures. The purpose here is to evaluate the feasibility of implementing landslide susceptibility analysis to identify regions that are more likely to suffer landslides. The analyses conducted in this study included in evaluating three different components based on the following effects: (1) creating three different datasets (herein referred as scenarios) based on four different feature selection methods (these datasets serve as landslide conditioning factors), (2) different ways of coding the datasets (observation types) for the analyses (raster versus vector approaches), and (3) different ways of conducting the susceptibility analyses based on three different algorithms (Random Forest, C4.5, and C5.0). In each scenario, thirteen different features with topographical, geologic, geotechnical, and hydrological features were included. In Scenario 1, the dataset was reduced by eliminating the least important factors/features (as identified by Information Gain, Chi-Square, and Gain Ratio measures). In Scenario 2, the dataset was also reduced but this time only the most important factors were selected (as identified by Random Forest measure). In Scenario 3, the data set included all available factors/features for the region (no feature selection method was applied to reduce the data). The study area for the LSA were kept the same as the one used for ALDM analyses, therefore the areas where landslides and non-landslides existed within the region were known. Thirty percent of the PDCNR data was used to train the machine learning algorithms to familiarize them for the common features observed in the areas of landslide and non-landslides. Once training was completed, the LSA analyses were conducted to test the validity of the proposed method. The remainder of the PDCNR data was used to check the accuracy of the LSA analyses conducted in this study. The results showed that the best prediction was obtained (where LSA outlines areas within the study area as most susceptible to landslide and the PDCNR data shows existence of landslides within that area) when the analyses were conducted with the Scenario 3 dataset that was coded as raster data and analyzed with C4.5 algorithm. Most studies in published literature choose the Random Forest algorithm to conduct landslide susceptibility analyses and model their data as vectors. This study showed that modeling the dataset as raster and using a less commonly considered algorithm provides better results. It presents a methodology on how the data can be compiled, modeled, and analyzed as well as within the scenarios evaluated, the most accurate algorithm to be used. In the ALDM analyses, the results were only based on geometrical features, whereas in the LSA analyses, the algorithms outline the areas within the region for landslide susceptibility. The zones that overlap by the ALDM and LSA should be treated as the most important areas within the region for the next landslides to trigger. Findings from this part of the study are used to study research questions 2 and 3. Details of this part of the study have been presented in Chapter 3.

Predicting the occurrence of the new landslides has always been a big challenge (if not impossible). **In the third part of the research**, the study focuses on developing an infiltration model for the study area (same as used in the first and second parts of the study) to determine the depth of rainfall penetrating into the ground. Although there could be many other triggering factors for a given landslide, in the study rainfall was chosen because it is regularly forecasted by the media and is widely available. The analyses conducted in this part of the study was based on the assumption that 100% of the rainwater infiltrates into the ground and depending on the intensity, the depth of infiltration increases. As the water infiltrates into the ground, it changes the geotechnical properties of the soil within the ground in terms of density and shear strength. These two properties are known to be most important for evaluating a specific slope for instabilities. Based on these changed soil properties, a new dataset of the region was created, modeled as raster data, and evaluated based on C4.5 decision tree algorithm (the algorithm that was previously defined in this study as the most effective/accurate one). The analyses were performed to capture three different rainfall intensity (high, moderate, and no rainfall). The results obtained from the no rainfall analyses were used to evaluate the potential accuracy of the model as with no rainfall, no soil properties were changed, and therefore no areas should be identified to trigger future landslides. This evaluation indicated that the conducted analyses had a ±2.5% accuracy. The analyses showed that the percentage of triggering landslides in the future is directly related to the intensity of the rainfall events. Because the focus of this last part of the study was to make predictions for the future, it was not possible to ground truth the predictions within the duration of this study. However, the methodology presented in this study provides an opportunity to create an early warning system for regions that have concerns regarding landslides. Even if the predictions may not be 100% accurate, it could be used to flag certain regions, which may save lives. Findings from this part of the study answered research question 4. Details of this part of the study have been presented in Chapter 4.

Study Area – Same for All Parts of the Study

There have been numerous studies that focuses on the landslide features in the west coast less studies were conducted on the east coast. Majority of the slides in the west coast are deep seated and the elevation differences between the toe and scarp are high. This allows capturing the location of the landslides much easier. However, in the east coast, majority of the slides are shallow (mostly through colluvium soil layers), and much less defined in terms of the elevation height differences between the toe and scarp. Therefore, to be able to better contribute to the existing knowledge, a study area in the east coast was selected. Detailed literature review indicated that the most active landslide areas within the east coast of the U.S. are within Commonwealth of Pennsylvania. However, Pennsylvania covers an area of approximately 45,000 square miles (117,000 square kilometers). Therefore, before selecting a specific region within Pennsylvania, an in-person meeting with the Pennsylvania Department of Conservation and Natural Resources (PADCNR) members took place in Harrisburg, PA. PADCNR team who has worked extensively over the years to map out the existing landslides provided insight information regarding the actual lay of the land and the extent of the publicly available dataset. Based on these considerations, the study area was selected as the Mansfield region of Pennsylvania (Fig. 1.1).

The Mansfield region not only covers areas of landslide and non-landslides but also has been extensively surveyed by LiDAR, which allowed the creation of a bare earth DEM for the study. Furthermore, PADCNR has an available dataset for the region showing the locations of the existing landslides that have been determined from ground observations and aerial photography. Therefore, the same area could be used for all three parts of the study.

The Mansfield region is located within the Tioga County of the Williamsport quadrangle, which contains major portions of the geological formations such as Deep Valleys, Glaciated High Plateau, and Glaciated Low Plateau sections of the Appalachian Mountains. Tioga County is one of the landslide-dense regions within the Mansfield region (approximately 135 square kilometers with 67 pre-identified landslides). These landslides are also relatively well-spaced from each other, providing the means to analyze wide spectrum of different cases as the site contains slump (rotational slides) and composite landslides (Delano and Wilshusen, 1999). The boundaries of the area are defined by the longitudes of -77.008657 to -77.185069 and latitudes of 41.769983 to 41.855648. The altitude within the area ranges from 329 to 565 m above mean sea level and decreases from west to east, where slope angles range between 0 and 69 degrees. The average slope in the area has been defined as approximately 35 degrees, which resembles an approximately 1 vertical to 1.4 horizontal slopes.

Slumps in the study have been defined by Delano and Wilshusen (1991) as rotational slides where the surface of rupture is concave-up. The head of the slump typically tilts back into the slope and the toe characteristically rises. Individual slumps commonly form segments of larger composite landslides and may exist by themselves as small discrete slope failures along a hillside. The slumps in the study area have been predominantly identified to occur in uniform fine-grained soil layers. The composite landslides in the study area are characterized by combination of different landslide types such as debris slides and slump or debris falls and debris flow.

An inventory map of landslides within the study area was created by PADCNR. Segments of the inventoried sections are shown in Fig. 1.2a. For the purposes of this study, the information provided by the PADCNR has been digitized and used for comparison of landslide mapping (Fig. 1.2b). Based on the inventory conducted by PADCNR, of the 47 areas of the 67 (70.1%) landslides are considered as slumps and 20 areas (29.9%) are composites. It is also important to mention that 55.2% of the existing landslides are considered historic while 44.8% of them have been listed to occur in early 1991 (Delano and Wilshusen, 1991). The active or recent landslides show clear, fresh scarps, distributed vegetation, fresh deposits at the toe, or other evidence of recent movement (Delano and Wilshusen, 1991). In order to capture the variety of the landslide regions within the study area, three different DEM tiles have been created. This was important to demonstrate that the proposed model could work in areas with abundance of landslides and with no landslides. Fig. 1.2b shows the created DEM tiles within the study area. These tiles were obtained from the PADCNR's data repository where each tile is referred as 59002140, 60002140, and 60002160. In order they are listed herein, each tile consisted of 15 (abundant landslide areas), 8 (Medium-range landslide areas), and 0 (no landslide) landslide areas in a given region respectively. Variation of landslide distributions and types within the selected area allowed a spectrum of terrain to be used in validation of the developed method.

Figures

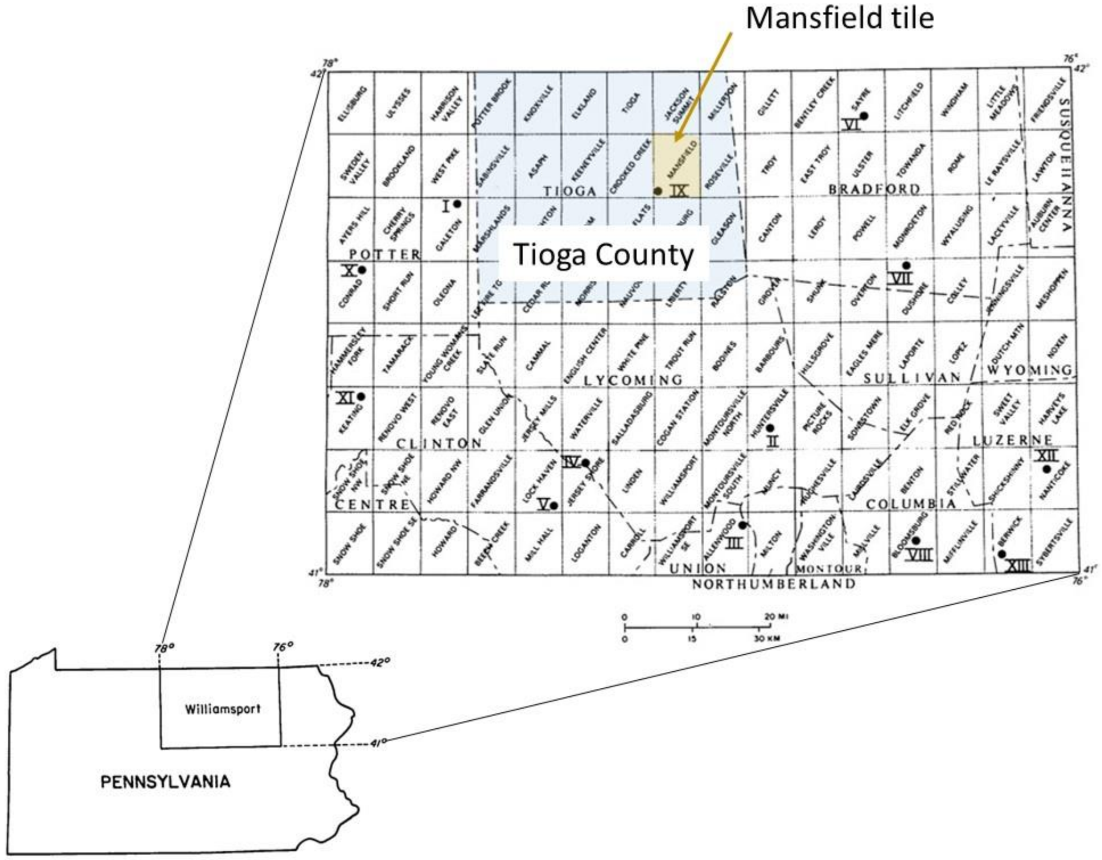


Figure 1.1. Location of the study area used in this study (adapted from Delano and Wilshusen, 1999)

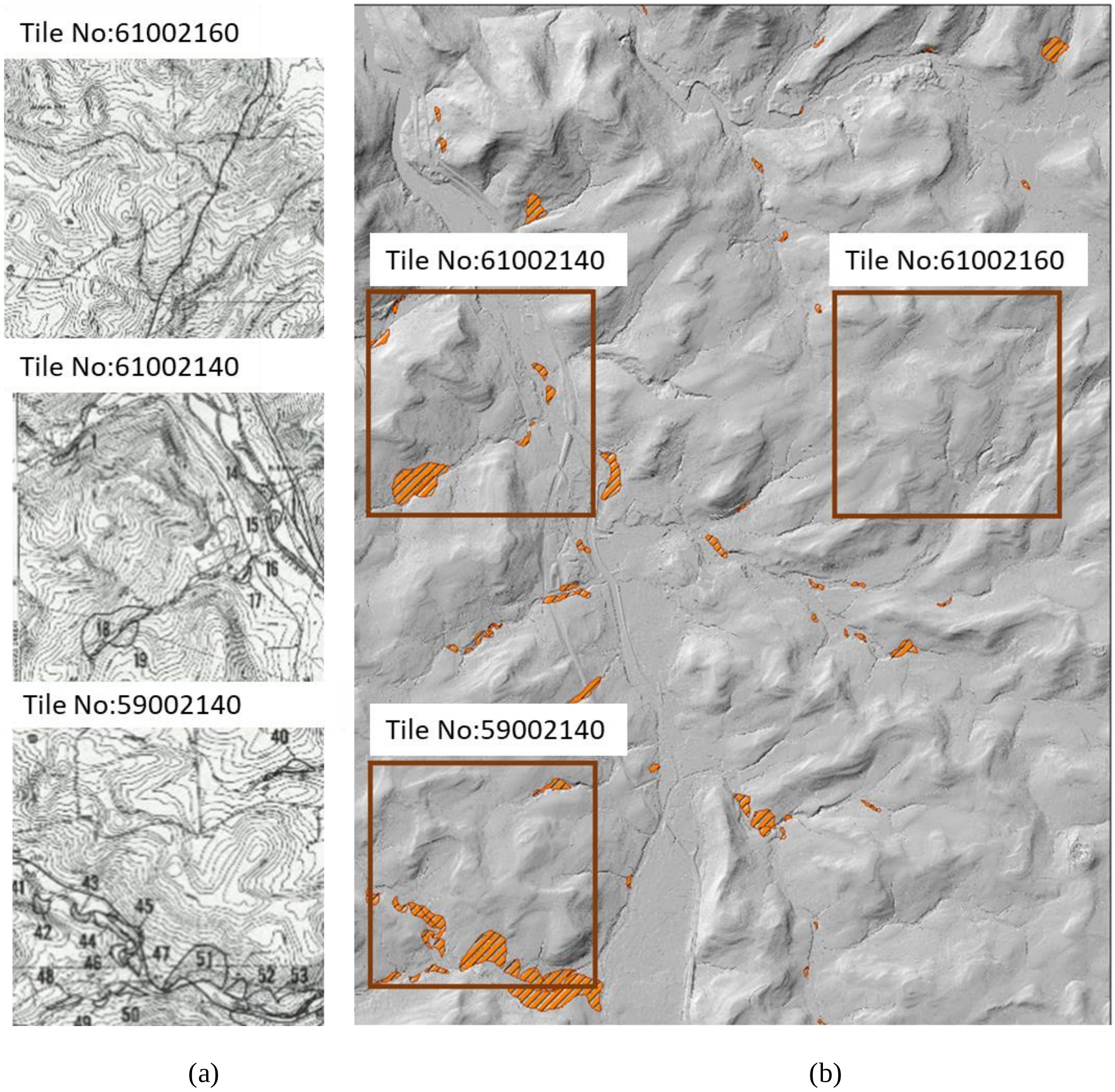


Fig.1.2. Locations of existing landslides within Mansfield area used in this study as shown in (a) PADCNR locations that have been defined by numbers (Delano and Wilshusen, 1999) and (b) digitized version on DEM map

CHAPTER 2: Automated Landslide Detection Model to Delineate the Extent of Existing Landslides[1]

Introduction

Landslide inventory analysis is the first and foremost step in landslide studies. The success of susceptibility and hazard analyses lies within the proper identification of the existing landslides. There are several recently-proposed methodologies to delineate landslides (Guzzetti et al., 2012; Bolstad, 2012; Pardeshi et al., 2013). The conventional approach for landslide inventorying involves using the data obtained through the field survey mapping and historical records, which is tedious to compile and difficult especially in heavily vegetated terrains (Pardeshi et al., 2013). Major challenges associated with conducting field surveys include the size of the landslide (often too large to be seen completely in the field), the viewpoint of the investigator (often inadequate to see all parts of a landslide with the same detail), and the fact that old landslides are often partially- or completely-covered with vegetation, or have been partly dismantled by other landslides, erosion processes, and human actions (Guzzetti et al., 2012). Visual interpretation and analysis of aerial photographs also has its own challenges as this approach requires experience, training, a systematic methodology, and well-defined interpretation that relies on the expert's opinion (Guzzetti et al., 2012). Although the traditional means of field

[1] The research described in this chapter has been submitted as a journal manuscript in February 2020.

survey is critically important, the availability of new remote technologies and satellite data offers an important opportunity to improve landslide inventory mapping.

One of the previously used methods in landslide inventory analysis is the interpretation of airborne or satellite imagery (Mwaniki et al., 2017). This interpretation in previous studies has been conducted by utilizing the geographical information systems (GIS) (Barlow et al., 2003; Colombo et al., 2005; Bolstad, 2012). However, there are challenges with using satellite imagery as this approach requires the users to specifically define the geometrical features seen in the image (implementation of supervised method) and to interpret the observed features even in areas with heavy vegetation cover. It is worth noting that the supervised classification requires predetermined specifications, defined by the operator in each area. Therefore, it is not an entirely automated means of mapping landslide features in a given area. On the other hand, the main technique used in interpreting satellite imagery involves in recognizing the difference in reflectance and shape of multiple colors. This technique sometimes causes significant errors since many existing landslides are covered by dense vegetation, and it is difficult to detect them by utilizing satellite imagery. Therefore, vegetated, older, dormant slides with subdued topography may be unrecognizable from air photos or multispectral digital imagery (McKean & Roering, 2004).

The use of light detection and ranging (LiDAR) data to create bare earth digital elevation models (DEM) is an effective means of mapping existing landslides in a given region. LiDAR works by emitting laser pulses at defined, horizontal and vertical angular increments to produce a 3D point cloud, containing XYZ coordinates that return a portion of the light pulse within range of the sensor (Leshchinsky et al. 2015). Various studies have utilized LiDAR data to evaluate topographical properties of landslide-prone terrain such as surface roughness, slope variance, and fractural dimension (Barlow et al., 2003; McKean and Roering, 2004; Glenn et al., 2006; Booth et al., 2009; Leshchinsky et al. 2015; Li et al., 2016). These studies demonstrate the ability to detect existing landslides with well-defined features. However, in regions such as east coast of the U.S., landslides are often smaller in scale and sporadic; further, their features such as scarp and toe tend to be not as well-defined. In these conditions, the previously determined landslide detection algorithms are not as effective because they require supervision wherein the user is expected to define features of the landslides (e.g., the height of the scarp, the slope of the body, and the toe of the landslide) in a given region. Therefore, without the proper training or site-specific knowledge about the study area, accurate delineation of the existing landslides becomes challenging and inconsistent.

In this study, a methodology to quickly identify the locations of the existing landslides using readily-available GIS tools and image processing algorithms without the need for any supervision from the user is presented. The presented methodology is evaluated by comparing the results against the ground truth defined by the trained landslide experts.

Development of Automated Landslide Detection Model (ALDM)

The automated landslide detection model (ALDM) is based on assessing a selected area to identify surficial geometric properties that characterize the roughness of the surface associated with recent landslides. The steps used to identify landslides are presented in Fig. 2.1. First, a bare earth digital elevation model must be used to identify the locations and features of the existing slopes within the selected areas. Using the focal statistics tool (Bolstad, 2012), the roughness of the terrain was quantified and converted into a binarized form. This information provided a quantitative means to evaluate whether the surface was smooth or hummocky, which was used to identify the location of the existing landslides. Unlike previous versions of the existing landslide delineation methods, the method described herein does not require any pre-defined landslide characteristics by expertise. However, as in all tools, the validation of the identified landslides must be confirmed by the ground truth defined by the trained landslide experts. Therefore, in this study such validation has also been performed to present the accuracy of the presented methodology.

Step 1: LiDAR data acquisition and development of DEM:

The first step of the ALDM requires the user to obtain a DEM of bare-earth surface elevation created from representative LiDAR points. In this study, a 1 m^2 pixel resolution DEM for the study area was obtained from PADCNR's map repository (PADCNR, 2019). Each pixel within the DEM represents an interpolated elevation between contours that were defined in a contour map. The DEM data of the study area is shown in Fig. 2.2.

In order to increase the speed of analyses, the DEM was reclassified to a 6 m^2 pixel resolution. The sensitivity of the results to the revised resolution was not investigated but the accuracy of the results with the reclassified resolution has been evaluated based on the landslides that occur both on east and west coast of the U.S. This evaluation was conducted to confirm the suitability of the proposed method to apply to the landslide features that are very different from each other (i.e. west coast landslides are typically deep-seated with well-defined scarps and the landslides in the east coast are shallow without as well-defined features).

Step 2: Slope Analyses

For each pixel in the DEM, slope was calculated based on finite difference approach, which is a standard algorithm tool embedded in most GIS platforms (Burrough et al., 2015). In this study, the angle of the slope of each neighborhood is computed using a 3 by 3pixel square. This is the smallest (highest resolution) option to compute slope angles as the process requires the pixel in the middle to be surrounded by equal number of pixels

Step 3: Roughness Analyses

After performing the slope analyses, the DEM is then analyzed using the Focal Statistics tool (Bolstad, 2012) of the GIS to determine the roughness. The output raster from roughness analyses is the result of a statistical function in a specific shape of neighborhood defined by the user. These analyses are performed to calculate the statistical value for each pixel within a specific neighborhood to determine the roughness of each pixel. The concept relies on the common knowledge of landslide analyses, where the older landslides would typically have a smoother surficial features and newer landslides have rougher, more well-defined features.

In Focal Statistics analyses, the neighborhoods can overlap with the neighborhood of another processing pixel. The size and type of this processing neighborhood can be selected as a factor of annulus, circle, rectangular, wedge, and user-defined shape analyses. In this study, 136 analyses were performed with annulus, circular, rectangular, and wedge shapes that consisted of 3, 5, and 7-pixel sizes. The most accurate results as defined by texture segmentation (the next step in these analyses) were obtained based on circular shape neighborhoods that consisted of 7 pixels. Fig. 2.3a depicts this configuration. In addition to the shape of the neighborhood, the Focal Statistics analyses require the user to define the function of the analyses, where the datasets are evaluated based on Majority, Maximum, Mean, Median, Range, and Variety functions. In this study, the most accurate results were obtained when the datasets were evaluated based on the "variety function". This function calculates the variety (the number of unique values) of the cells in the neighborhood. For instance, if there are seven unique values (cells) in the neighborhood of a processing cell, the value for this cell would be seven (Fig. 2.3b).

Step 4: Texture Segmentation Analyses

Texture segmentation is a process of classifying multiple sets of pixels into segments to make imagery more meaningful and easier to analyze. Based on Malik et al. (2001), image segmentation can be classified into two broad categories as region-based and contour-based approaches. In region-based approach, image properties such as brightness, color, and texture are used to define partitions of pixels. Contour-based approach starts with edge detection to catch the contours of different partitions (as commonly used in facial recognition tools). In this study, the texture segmentation is conducted with region-based Image Binarizing technique (Mathworks, 2019) that is an available tool in MATLAB software.

The tool is used to first convert the results from Focal Statistics (that comes as a color map) into a grayscale image. The converted image is then processed to develop a binary image based on luminance threshold. This is achieved by replacing all pixels in the input image with a number of 0 or 1. In this method, the value of 1 represents the white color pixels (which indicated the location of the landslide features) and 0 represents the different shades of gray scale (indicating the location of non-landslide features). The luminance threshold for this study was selected to be 0.65 indicating that the pixels with the values below this threshold is depicted as black and above as white. After the binarizing (Mathworks, 2019) step, the texture segmentation analyses is continued based on a function that removes small features from the binary image (in Mathlab software this function is referred as bwareopen (Mathworks, 2019)). Each tile selected in this study approximately consisted of 9,765,625 pixels. As a result of bwareopen function, the features shown on the binary image with less than 50 pixels were removed. This step helped removing what may be considered as a noise in defining the landslide and non-landslide features. These features were then used to create the actual areas, which is obtained by using imclose function of the Mathlab image processing toolbox (Mathworks, 2019), where the open circles are closed.

Methodology Followed on Validation of the Developed ALDM

To be able to validate the results obtained from ALDM, landslide data directly obtained from PADCNR has been compared against detections obtained from the model developed in this study. Additionally, based on the Hill Shade maps of the study area, the authors have visually depicted the locations of the existing landslides manually. This approach herein is referred as Visual Detection (VD) analyses. Fig. 2.2 shows the transition of the DEM data (Fig. 2.2a) within the study area to Hill shade map (Fig. 2.2b) that is created by utilizing ArcGIS. LiDAR data from 2008 has been utilized to create the Hill shade map for this area. Comparison of the Hill shade map (Fig. 2.2b) with the ortho-photograph of the area (Fig. 2.2c) shows that the geometrical features of the study area.

Overall, the analyses performed for validation were conducted based on the Set Theory and Venn diagram that allows the users to compare all possible logical relationships between the existence and non-existence of landslides in a given region. A Venn diagram consists of multiple overlapping closed circles each representing a data set (Fig. 2.4). The interior part of the circle symbolically represents the elements of the set (i.e., circles A and B in Fig. 2.4), while the exterior part represents elements that are not members of the given set (i.e., area D in Fig. 2.4). The Intersection of two circles (1 ∩ 2) is the area covered or the shared results between two sets (i.e., area C in Fig. 2.4). Based on this information, a confusion matrix may be created where accuracy of the predictions can be validated. Data that falls within the area of false negative (FN) would indicate that although a landslide exists, the model captures it as a non-landslide area. False positive (FP) would mean the opposite of FN where a non-landslide area is incorrectly captured as a landslide area. High values of true predictions (positive – TP or negative TN) would indicate accurate results meaning both the landslides and non-landslides are predicted correctly.

ALDM Results Obtained within the Study Area and Validation

ALDM analyses were conducted with three tiles (i.e., one with abundant landslide areas, one with modest landslide areas, and one with no landslide areas). The results obtained from each tile have been compared against the PADCNR field data and VD to check the validation. Below, describes the results and validation comparisons for each tile used in this study.

Abundant landslide areas: Fig. 2.5a shows the detection of the landslides in the area that is referred as "abundant landslide areas - tile no. 59002140" from the ALDM developed in this study. Figs. 2.5b and 2.5c show the results obtained from PADCNR and VD analyses respectively. Comparison of the ALDM results and the data for the validation has been gathered for a visual presentation in Fig. 2.5d. Using GIS as a tool, PADCNR and VD data have separately been compared against the detection results from ALDM and the outcome of these comparisons have been tabulated in Table 2.1. In this comparison, the TPR value (as defined in Fig. 2.4) is considered as the indicator of the accuracy to capture the existence of the landslides within the area. The results show that when the ALDM is compared against the PADCNR data, the accuracy of the predictions is within 74%. Although there is no threshold value that defines an acceptable range of accuracy in these types of analyses, when the percentages of true landslide locations predicted by ALDM are compared with other types of automated methods, 74% is considered as very accurate

(Leshchinsky et al., 2015). In the same analyses, the accuracy of predicting non-landslide areas was even higher with the TNR values (as defined in Fig. 2.4) being in the range of 93%. When the ALDM results were compared against the VD data, the accuracy of the predictions was also high. This is because the VD data at this location was similar to those captured by the original PADCNR survey data.

Modest landslide areas: A similar figure as Fig. 2.5 has not been presented herein but following the same technique, the ALDM data from the tile number 61002140 have been compared against the PADCNR inventory data. The overall accuracy of ALDM obtained from this area was similar to the results obtained from the area with abundant landslides (i.e., TNR values of 95% vs. 93% and TPR values of 78% vs. 74% respectively). However, when the same comparison was made based on VD data, the accuracy of the predictions was slightly higher for the non-landslide areas but lower for landslide areas. Although it should be noted that the comparison of the ALDM and VD results is qualitative because the accuracy of the VD data highly depends on the interpretation of the person who evaluates the features of the Hill shade map. This comparison was conducted because VD analyses is one of the viable tools that are utilized by the DCNR agencies to assess the delineation of landslides from LiDAR data (Delano and Wilshusen, 1999).

Non-landslide areas: Tile number 61002160 was used to perform the ALDM analyses in an area with no landslides to evaluate the performance of the predictions when no landslides should be detected. Since there were no landslide data available at this location, the accuracy of the predictions of non-landslide areas were primarily based on the TNR values. The comparison of the ALDM and PADCNR data shows the TNR values being in the order of 99%, which shows agreement with what is predicted and observed in the field. As for the other locations, when the ALDM was also compared against the VD data, however the accuracy of this comparison was not as high as was for the PADCNR data. This is because during the VD, the authors have interpreted some of the areas as landslides but the computer-generated interpretation via ALDM did not capture those features.

Testing of the ALDM Detection Results from Areas Outside of the Study Area: Oso Region of the Washington State

The results shown in Table 2.1 presented the validity of the method described in this study. However, to test the performance of the ALDM in other parts of the world, where the type and size of the landslides may differ than the ones evaluated in this study, testing of the model has to be checked. To achieve this goal, predictions obtained from the ALDM method described in this study was compared with the landslide that has occurred in March 2014 at the infamous Oso landslide in Washington State. This area was particularly selected because there are several landslide areas identified by USGS in this area with significant age differences (i.e., ranging from as old as 14,000 years to the youngest era between 2006 and 2014) (Haugerud, 2014) and the size and shape of those landslides also significantly differ from the ones evaluated at the study area.

The results of the ALDM analyses in the Oso landslide area are shown in Fig. 2.6, which is projected over the Hill shade map created from United States Geological Survey (USGS) dataset (Haugerud, 2014). When compared, the area identified by the ALDM very closely match with the area identified by USGS where the 2014 Oso landslide has occurred.

However, the area that is captured by the ALDM does not cover the entire area that is defined by USGS as the footprint of the Oso landslide. Also, the previously defined landslides within the region have not been captured by the ALDM (Fig. 2.6). This shows that the selection of roughness analyses and the luminance thresholds have an affect over the results. The results shown in Fig. 2.6 are based on the parameters that were determined from the analyses performed for the study area (Fig. 1.2). These parameters were based on the size and shape of the neighborhood as circular with 7 pixels (as shown in Fig. 2.3a), the function of the roughness analyses based on the feature referred in ArcGIS as "variety function" (Fig. 2.3b), and the luminance threshold as 0.65. The effects of the age of the landslide on the computer-generated predictions have previously been discussed by Leshchinsky et al. (2015), where the older landslides show smoother surfaces and the younger ones contain more hummocky areas.

The results shown in Fig. 2.6 indicate that the ALDM analyses performed in this study was accurate in predicting the hummocky areas (which is the area of the newest landslide) but not so much in predicting the areas with the smoother surfaces that are associated with the landslides that occurred thousands of years ago. To demonstrate the practical implications of the ALDM method, the luminance thresholds of the roughness analyses that were previously used for the area where the Oso landslide has occurred were adjusted. Fig. 2.7 shows the outcome of these analyses where the size and shape of the neighborhood was selected as circular 5 pixels, the function of the roughness analyses was based on range, and the luminance threshold was 0.5. The results show a good agreement between what would be the visually detected landslides and the landslides detected by

ALDM. It is claimed that once the user fine tunes these adjustments, the ALDM method described in this study could be used in any parts of the world to depict the locations of the existing landslides. However, it is important to note that the users of such automated landslide detection methods must understand the specific features exist in landslide areas. The advantage of the method described in this study compared to the other existing computer-based methods is that, the effort required to fine tune the ALDM is very minimal and very straightforward and all of the tools are readily available through ArcGIS, MATLAB image processing toolbox, and existing LiDAR data. The method proposed does not need supervision, meaning once the model is fine-tuned to a region, other adjustments are not required to obtain reasonably high accurate results (as shown in Table 2.1 and Fig. 2.7).

Conclusions

This study demonstrated the use of the automated landslide detection model (ALDM) that utilizes the DEM created from LiDAR based data, readily available GIS tools, MATLAB software, and statistical validation methods. The concept of the ALDM is to identify the roughness of the slope surfaces and use the image processing techniques to delineate the extent of the existing landslides. It is developed as a simple but effective tool that is based on readily available software packages. The analyses require the user to initially set thresholds for roughness and texture segmentation analyses, however once these thresholds are established, the analyses are conducted in large areas without supervision and the need to re-define these thresholds. However, if the characteristics of the landslides from one region to another varies significantly (for example shallow landslides vs. deep seated landslides), revisions to the thresholds might be necessary as demonstrated in this study.

Three areas with LiDAR bare earth digital elevation models (DEMs) have been used to test the proposed approach, each consisting of a varying range of mapped landslide features in Pennsylvania. The results obtained were compared against data from PDCNR and landslides that were determined visually from the Hill shade map (a technique that is implemented by some of the DCNR agencies to delineate landslides). The results demonstrate that the proposed ALDM method was able to accurately capture both the landslides and non-landslides in all of the areas evaluated with accuracies of 70% and 92% respectively. This allows the users to evaluate large areas with minimal effort in very short time as compared to visual detection methods.

Additionally, the study also showed that the proposed ALDM method could be implemented in regions where the landslide sizes and features could be significantly different. To present this demonstration, several landslides that have occurred in the west coast of the U.S. at significant time differences have been selected. The results showed that even if the analyses are conducted with the pre-defined thresholds from other regions (i.e., the thresholds defined from the analyses conducted with landslides in PA) the remnants of the landslides are detected. However, more accurate delineation requires these thresholds to be re-defined. The results also showed that although the ALDM was effective in capturing the younger landslides (i.e., less than 25 years), the method is not as effective in capturing the older landslides (landslides that have occurred 500 years or older). This is because although the thresholds could be redefined, the texture segmentation analysis has a limitation. Considering that in most cases the real dangers of landslides are associated with younger landslides that can remobilize, the proposed method is a viable technique to delineate landslides.

Tables

Table 2.1. Comparison of the Results Obtained from ALDM, PADCNR and VD Methods

Tile No.	Area (m^2)	Mode	TNR	FPR	TPR	FNR
59002140[1]	9289996.30	ALDM vs. PADCNR	92.55%	7.45%	74.03%	25.97%
		ALDM vs. VD	94.35%	5.65%	86.88%	13.12%
61002140[2]	9290340.57	ALDM vs. PADCNR	94.63%	5.37%	78.07%	21.93%
		ALDM vs. VD	96.26%	3.74%	68.69%	31.31%
61002160[3]	9290315.73	ALDM vs. PADCNR	99.20%	0.80%	N/A	N/A
		ALDM vs. VD	99.73%	0.27%	20.06%	79.94%

Notes: ALDM: Automated landside detection method
PADCNR: Field surveyed landslide locations
VD: Landslide locations identified by the authors based on visual detection
N/A: Not applicable
[1] Area with abundant landslides
[2] Area with moderate abundant landslides
[3] Area with no landslides

Figures

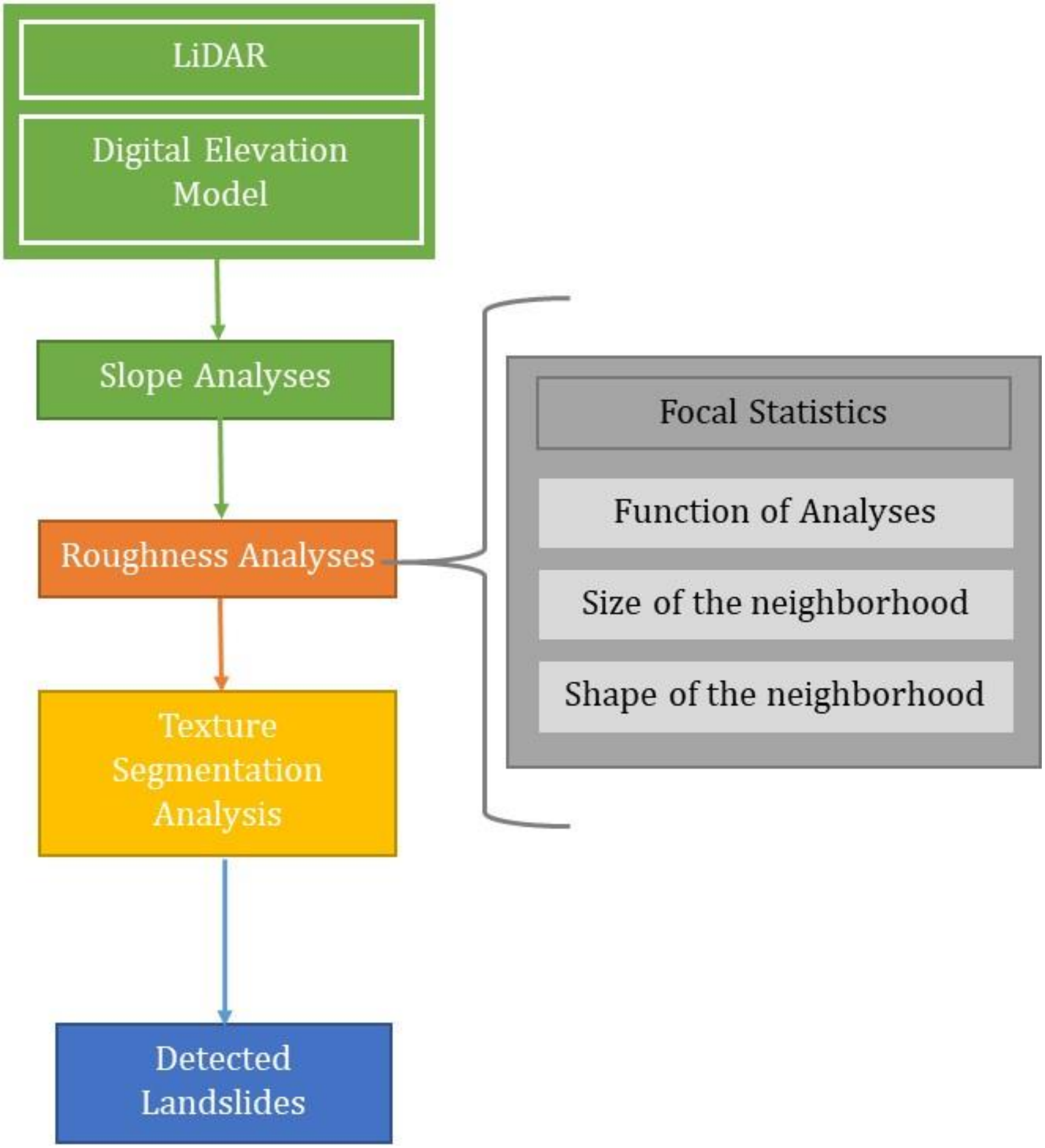


Figure 2.1. Overall steps used to develop automated landslide detection model

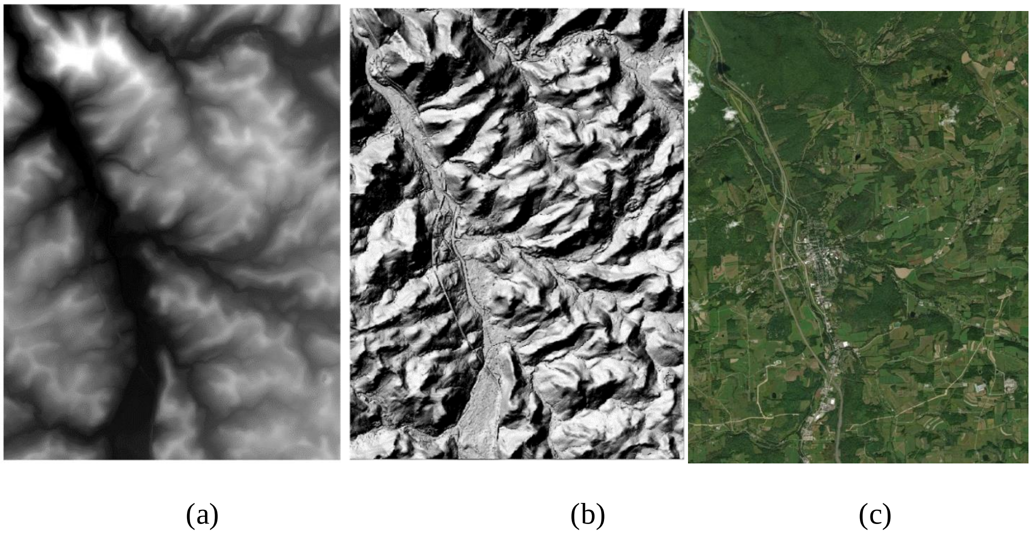
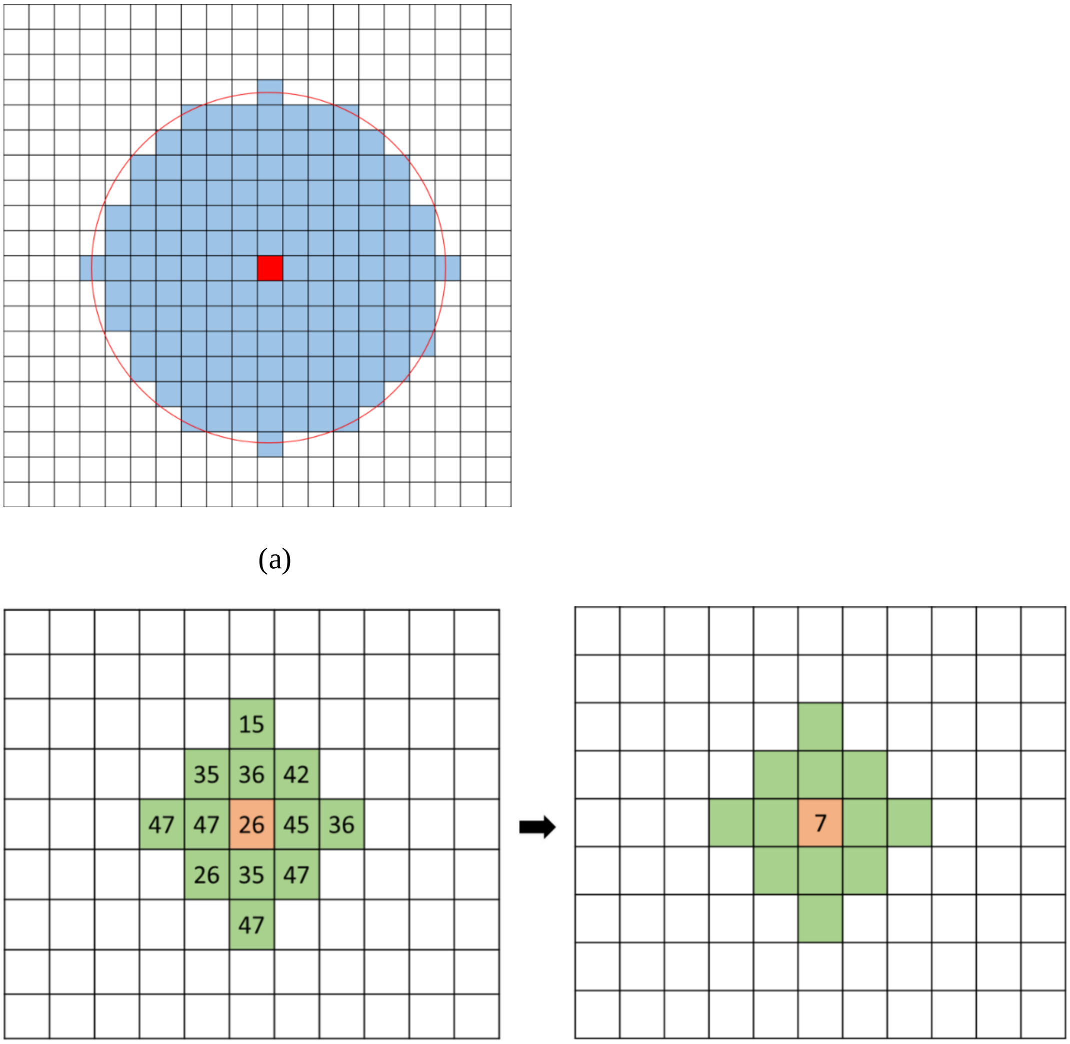


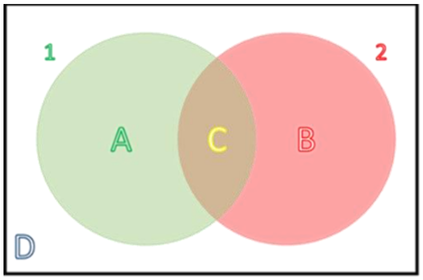
Fig. 2.2. PADCNR data showing (a) DEM created from LiDAR data of the study area, (b) Hill shade model of the DEM from LiDAR, and (c) ortho-photograph of the area shown in DEM.

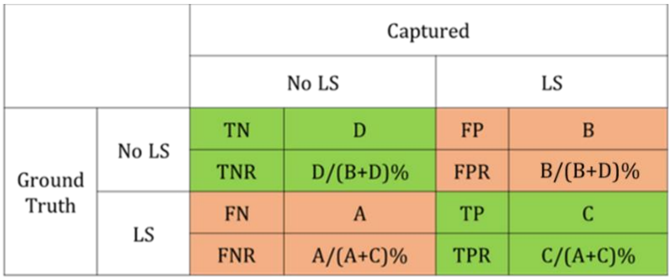


(a)

(b)

Fig. 2.3. Pixel configurations used in this study for (a) 7-pixel circular neighborhood and (b) roughness analyses with variety function





		Captured			
		No LS		LS	
Ground Truth	No LS	TN	D	FP	B
		TNR	D/(B+D)%	FPR	B/(B+D)%
	LS	FN	A	TP	C
		FNR	A/(A+C)%	TPR	C/(A+C)%

Notes: Circle 1 - Ground truth based on known landslide locations

Circle 2 – Results from Automated Landslide Detection Model

Area shown in A - Represents False Negative (FN)

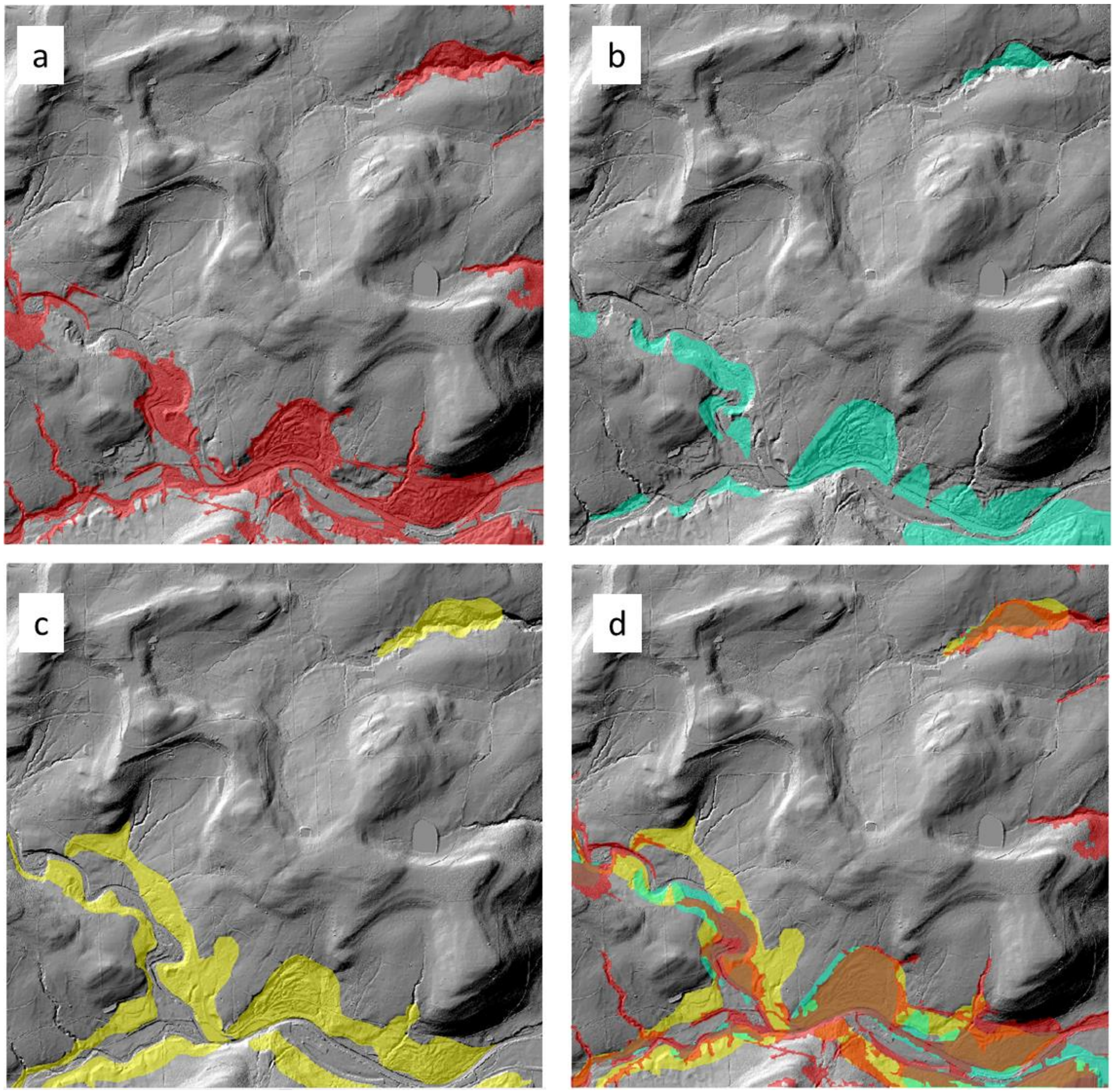
Area shown in B - Represents False Positive (FP)

Area shown in C - Represents True Positive (TP)

Area shown in D - Represents True Negative (TN)

(a) (b)

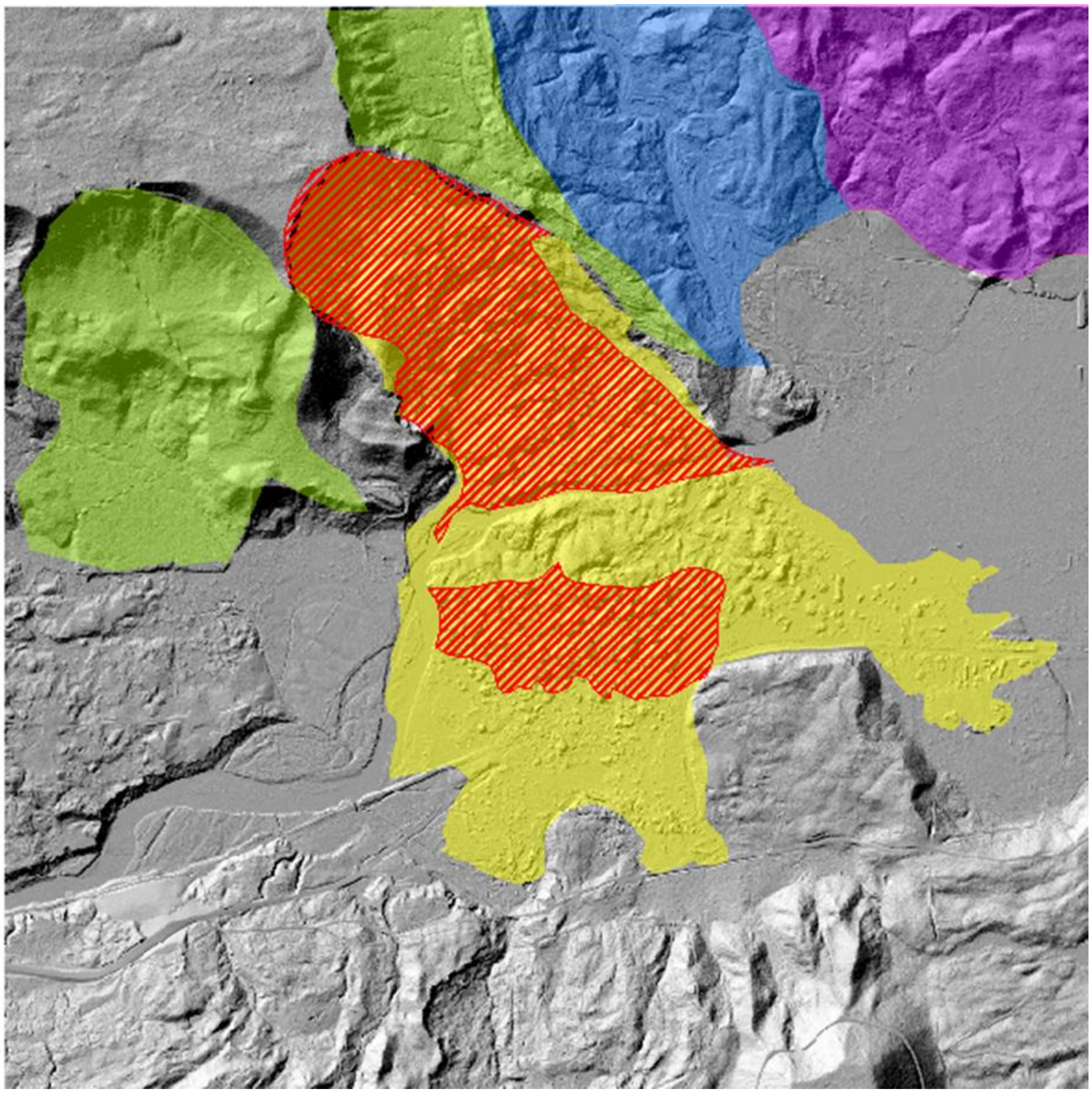
Fig. 2.4. Validation analyses based on (a) Venn diagram and (b) confusion matrix created from Venn diagram



Notes:

Red color: ALDM predictions
Turquoise color: PADCNR predictions
Yellow color: VD predictions

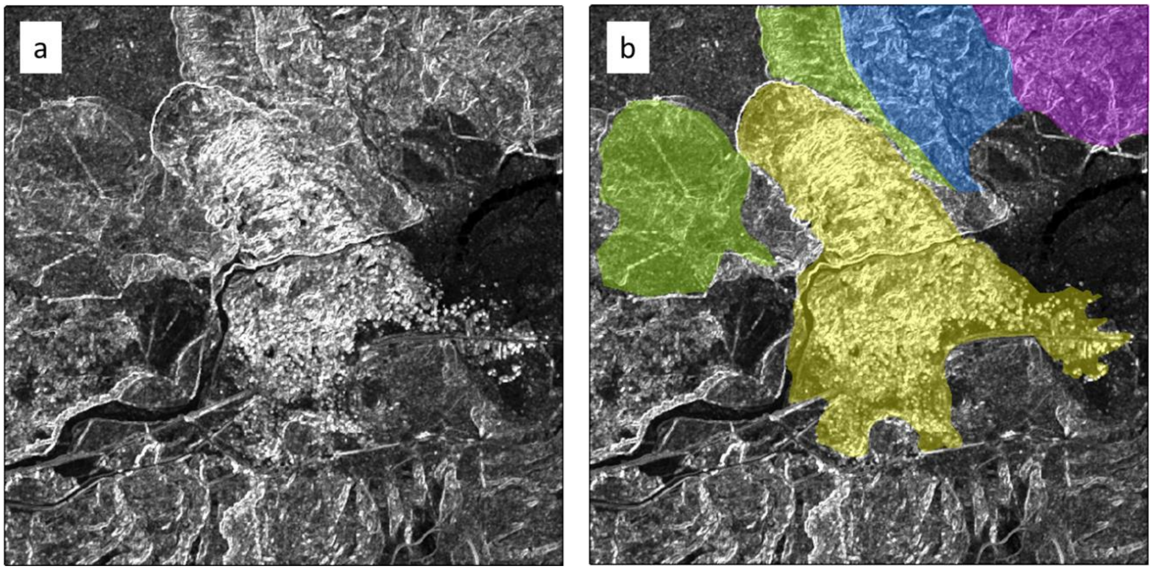
Fig. 2.5. Detections in area considered with abundant landslides based on (a) ALDM, (b) PADCNR, (c) VD, and (d) all three methods: ALDM, PADCNR, and VD.



Notes: The above figure has been re-created based on USGS identified landslides (Haugerud, 2014)

Red Hatched color: ALDM predictions obtained in this study
Yellow color: Oso Landslide mapped by USGS (most recent, landslide occurred in 2014)
Blue color: Pre-Oso Landslide, less than 500 years old mapped by USGS
Violet color: Pre-Oso Landslide, 500 – 2000 years old mapped by USGS
Green color: Pre-Oso Landslide, more than 5000 years old mapped by USGS

Fig. 2.6. Comparison of the ALDM results with landslides identified by USGS in Stillaguamish Valley in Washington.



Notes: Yellow color: Oso Landslide
Blue color: Pre-Oso Landslide, less than 500 years old
Violet color: Pre-Oso Landslide, 500 – 2000 years old
Green color: Pre-Oso Landslide, more than 5000 years old

Fig. 2.7. ALDM results of the Stillaguamish Valley in Washington (a) obtained after site specific revisions and (b) superimposed over the landslides delineated by VD method in this study.

CHAPTER 3: LANDSLIDE SUSCEPTIBILITY ANALYSES BASED ON RANDOM FOREST, C4.5, AND C5.0 ALGORITHMS USING BALANCED AND UNBALANCED DATASETS[2]

Introduction

The focus of this study was to develop a landslide susceptibility analysis (LSA) approach to evaluate the likelihood of landslide occurrence based on the conditioning (or preparatory) factors and to compare the performance of different LSA algorithms against multiple datasets and observation types (raster vs vector). Through the use of more accessible data and processing systems such as GIS (Geographical Information System), landslide hazard assessment and risk reduction can provide more useful and accurate information to the public and private sectors, governmental agencies, and the scientific community (Shahabi and Hashim, 2015). There are three components that determine the result of LSA: 1) features and dataset combination (Yalcin, 2011), 2) landslide observations (Trigila et al., 2015), and 3) algorithms used for evaluation (Alimohammadlou et al., 2014; Saito et al., 2009; Yeon et al., 2010). Change in these components may affect the accuracy of LSA or the ability of landslide forecasting. Accordingly, the main objective of this study was to evaluate the performance of three

[2] The research described in this chapter has been submitted as a journal manuscript at the beginning of April 2020.

different landslide susceptibility algorithms based on the variation in features and dataset combinations. Previous research on different features, and dataset combinations, as well as different landslide algorithms evaluated to date are summarized in the subsequent sections.

Features and dataset combinations

Critical in studying LSA is the evaluation of the most effective causes and triggers. Landslide process never occurs from a single cause, and to achieve an accurate understanding, a collection of parameters should be analyzed interactively. Many researchers have discussed these parameters. Crozier (1986) categorized landslide factors into preparatory (or conditioning) factors and triggering factors. Popescu (1994) divided the conditioning factors into: geomorphological processes, physical processes, and human influences. Wu and Sidle (1995) categorized the preparatory factors as geology, slope gradient and aspect, elevation, soil geotechnical properties, vegetation cover and long-term drainage patterns, and weathering. These views have been extended by Yalcin (2011), who described the conditioning factors as lithology, slope, aspect, elevation, vegetation cover, discontinuity, and the location of a nearby river or road.

In landslide susceptibility analysis, the quality of the evaluation depends on both the selected models and the quality of the input data. Ozdemir and Altural (2013) used the topographic factors (derived from a map based on Digital Elevation Model (DEM)), geology, land use/land cover, and precipitation data to create a landslide susceptibility mapping of Turkey. In another study, Pham et al. (2017b) assessed landslide susceptibility in India based on the topographic map, soil map, land cover map, and meteorological data.

In a study of landslide susceptibility analysis in the Shaanxi Province, China, Chen et al. (2017) used the topographic features of the slope, the lithology, land use, and the rainfall, to predict landslides. In general, the parameters used for landslide susceptibility analysis can be divided into categories, such as, geological, hydrological, topographical, and geotechnical features. However, all of the previously mentioned studies have mostly used geological and topographical parameters of the study area but have not included the geotechnical and hydrological parameters (Ozdemir and Altural, 2013; Chen et al., 2017). These studies would have been more accurate if they would have also included geotechnical parameters (such as soil engineering and index properties) and hydrological parameters (such as groundwater level, hydraulic conductivity, and soil moisture). Our study examines a larger set of landslides' contributing factors, including geological, geotechnical, and hydrological parameters.

Landslide Observations

The landslide inventory map is the first step in the analysis because it is believed that the past landslides are the indicators for the future events ((Guzzetti et al., 1999). Therefore, using an inventory map with detailed and accurate information will improve the accuracy of the results. In general, the inventory map and the existence of landslides in a given region is used a response/predictor in the landslide susceptibility model (Chen et al., 2017). In previous work, the observations are generally modeled as vector or point data on the area (Pham et al., 2017b). This means that in LSA model, each landslide occurrence is represented as a point on a map and in the dataset. The modeling of the landslide observations also captures the proportions of the landslides within an area. Most of the studies have used this data to select areas for their analyses where proportion of landslide and non-landslide areas are equal to each other ((Ozdemir & Altural, 2013). The primary reasons for this approach are due to the ease of the computational efforts and the sensitivity of the assessment methods to the unbalance datasets. One approach used by researchers to alleviate this limitation is by choosing random points on the source data (map) to create vector datasets ((Trigila et al., 2015), Hong et al., 2015, (Youssef et al., 2016) Pham et al., 2017b).

Algorithms

The landslide susceptibility models can be qualitative or quantitative. The qualitative approaches rely only on expert judgment and involve directly mapping the geomorphology to assess the susceptibility based on factors defined by the expert (Aleotti and Chowdhury, 1999). The quantitative approaches are primarily based on statistical and machine learning methods, such as Frequency Ratio (Ozdemir and Altural, 2013; Hong et al., 2015b), Logistic Regression (Oh et al., 2010; Ozdemir and Altural, 2013; Youssef et al., 2016), Random Forest (Trigila et al., 2015; Chen et al., 2017), Classification and Regression Trees (Chen et al., 2017; Wu et al., 2020), Artificial Neural Networks (Alimohammadlou et al., 2014; Bui et al., 2016; Pham et al., 2017a), and Support Vector Machines (Colkesen et al., 2016). Frequency Ratio (FR), Logistic Regression (LR), and Random Forest (RF) are ones of the most commonly considered methods.

The Frequency Ratio is a simple method of calculating the probabilistic relationship between dependent and independent variables such as multiple maps. In landslide susceptibility analysis, FR is the ratio of the area where landslides have occurred to the total study area and the ratio of the landslide occurrence probability to the non-occurrence for a given attribute (Ozdemir and Altural, 2013). Although it is an easy to use model, the FR analysis only estimates the performance of a single factor and does not produce the whole range of predictors.

The Logistic Regression method uses the values of a set of predictors identified within the landslide areas to predict susceptibility for landslides in other areas. Oh et al. (2010) show that this method has higher prediction accuracy than the frequency ratio and the artificial neural network methods. However, this method is not very suitable for unbalanced data where the number of non-landslide observations (pixels) is greater than the observations in the landslide area. The Random Forest method first classifies the characteristics of different areas to create a set of trees that are aggregated to compute landslide classifications. It then resamples this data and randomly replaces and changes the predictive set of variables over the different tree induction processes to create the landslide susceptibility model. The predictive variables used in this model can be based on numerical or categorical data. In recent years, Youssef et al. (2016) and Chen et al. (2017) are among the researchers who have implemented this method to predict landslide susceptibility in Saudi Arabia and China, respectively, with claimed range of success rates of 78% to 83%. They used vector and balanced datasets to perform their analyses.

Purpose of the Study for Landslide Susceptibility Analyses

The purpose of this study was to evaluate the effectiveness of the methods used to predict the susceptibility of an area to future landslides. The study included the evaluation of three different components of landslide susceptibility analysis: (1) input data associated with past landslide observations; (2) characterization of the study area based on geo-features such as geotechnical, geomorphological, and hydrological properties; and (3) suitability of performing the analysis with different data analyses algorithms. The importance of this study as it relates to each of these components is discussed below.

The previous literature predominantly indicates that susceptibility analyses were conducted with characterizing the input and output parameters as multiple vector layers representing different features (Ozdemir and Altural, 2013; Hong et al., 2015a; Trigila et al., 2015; Youssef et al., 2016; Chen et al., 2017; Shirzadi et al., 2019; Wu et al., 2020). This means, random points are selected from both landslide and non-landslide regions to represent the variables for each dataset. This approach is used frequently in regions where the amount of data is limited. In this study, a region with a more detailed data set was selected and the landslide and non-landslide areas were characterized based on both the vector and the raster datasets. Unlike the randomly chosen points, the raster dataset includes features associated with every point within a region. This allowed the study to compare the performance of LSA with vector versus raster datasets.

LSA is frequently conducted without considering the detailed geo-features of the area of interest, although it is known that specific geo-features play a major role in landslide occurrence. Geotechnical engineers typically deal with this problem by performing specific stability analyses on a very small area (measured in meters or feet), where the properties and dimensions are well-known. This approach is not suitable when the area of interest is a city or a large region (where the dimensions are within hundreds of square kilometers or miles). In this study, we integrated the important geo-features into the LSA by modeling the areas with datasets including digital elevation model, soil properties, and groundwater levels with depth. These datasets are then modeled both as vector and raster in the LSA. The study also focused on determining the importance of the specific geo-features by using four different feature selection models: information gain, chi-square, random forest, and gain ratio.

There are many different data analysis algorithms available in the literature. These algorithms may be used for many different purposes one of which being to perform LSA. In this study, three different algorithms, Random Forest (RF), C4.5, and C5.0 (which is the newer version of the C4.5), were evaluated to conduct LSA. The RF algorithm has been used in different LSA studies as a standard method to evaluate the results of landslide susceptibility analyses (Chen et al., 2017; Trigila et al., 2015; Youssef et al., 2016). In this study, the accuracy and error rates of the C4.5 and C5.0 methods have been compared to those of the RF algorithm. Overall, comparison of these three different approaches allowed the study to evaluate the performance of different algorithms to estimate the accuracy of the LSA model with different input parameters and their combinations. Both C4.5 and C5.0 were included in this study because they have different sensitivity to dataset balance (Saito et al., 2009). Although the findings of this study are associated with a specific dataset obtained from a given region, the methodology presented herein may be applied to other regions as well. This research identified the best approach to conduct LSA.

Components of the Model Built for This Research

Inventory Maps

Landslide inventory maps for the study area were obtained from the Pennsylvania Department of Conservation and Natural Resources (PADCNR). They were created based on the actual field surveys and visual interpretations conducted by PADCNR. The study area contains 67 landslide locations of different sizes and types. According to Delano and Wilshusen (1999), most of these landslides are considered active. The location of each landslide has been used in the model built for this research to differentiate between the properties of the regions with and without landslides.

Conditioning Factors

Conditioning factors refer to the properties of the area that are associated with landslides. As shown in the left-hand side of Fig. 3.1, in this study, the conditioning factors included the geological properties (i.e., type of bedrock at depth and at surface), hydrological properties (i.e., groundwater level, hydraulic conductivity of the ground, Skempton value – indicator of pore pressure in the ground), geotechnical properties (i.e., soil type, the percentage of the sand, silt, and clay size fractions within the soil, the density of the soil, and Atterberg limits), and geometrical properties (i.e., elevation of the ground as determined from the digital elevation map created in this study and slope percentage). When combined, these factors capture a comprehensive range of information that is critical for any slope instability evaluation. All of this information was obtained from PADCNR and the United States Department of Agriculture (USDA), being readily available for the region within the study area.

The data for the geotechnical/geological properties were provided in the database as a function of depth from the ground surface. For the purpose of this study, the conditioning factors were categorized for two different layers. The first layer is from the ground surface (0) to the depth of 30 centimeter (12 inches) which is also considered as the topsoil layer. The second layer is from the 30 cm to a depth until the bedrock is reached. This approach is justified by the fact that most of the landslides in this region were previously identified to be surficial, where the surface conditions of the slope impact the stability (Delano and Wilshusen, 1999).

The topographical parameters were derived from the DEM that was available through PADCNR to create parcels with 20 ft. × 20 ft size. Based on this DEM, using the ArcGIS platform, a map of slope angles has been created for the entire study area. The groundwater level in the study area has also been mapped out based on the static water levels (feet below the surface) provided by the Pennsylvania Spatial Data Access (PASDA) from each of the wells located within the study area. The static water level (feet below the ground surface) has been added to the ArcGIS as a vector layer with the location and depth of each well. This information was then used to create a contour map and raster layer of the groundwater level.

The study area has 15 different types of surficial geological features, which ranged from what is referred as Wisconsin, glacial, and re-sedimented tills to alluvium that contained stratified silt, sand, gravel, and boulder size particles. There were three different types of bedrock identified at the site consisting of Huntley Mountain Formation, Catskill Formation, and Lock Haven Formation. All the conditioning factors identified in this study and the location of the landslides from the inventory maps have been characterized as both vector and raster dataset. This is shown in the middle-top of Fig. 3.1, under Input.

Methods of Analyses

The combinations of all the conditioning factors have been used to create twenty-two layers of raster and vector data, which were utilized to perform the analyses. One of these layers was allocated to identify the areas with and without the landslides and the remaining twenty-one layers were developed as the predictors (or variables) for the analyses (see Fig. 3.1 – the variables column in the right side). Each raster layer was created from 3,645,234 pixels covering the entire study area, and each of the vector layers was created based on randomly chosen points from the dataset.

The analyses were performed by dividing the dataset into two groups, for training and testing, with a 70/30 ratio. In the first (training) group, 70% of the randomly selected data from landslide and non-landslide observations were evaluated as part of the analyses performed with the susceptibility algorithms (see the middle of Fig. 3.1). This information was used to train the model to predict the regions of landslides. The second (testing) group (30%) was used to test the effectiveness of the model (i.e., the output of the landslide susceptibility analyses – LSA, see Fig. 3.1). Unlike the previous studies, the research presented in this study has been verified not by generating random field/site conditions but by using the actual data that was directly obtained from the specific site that was used for the study. Therefore, this study not only presents the methodology for the analyses but also an actual evaluation of a real case study.

The dataset combinations (i.e., data generated both with vector and raster layers) have been evaluated by implementing three different scenarios, which involved utilizing four different feature selection methods (FSMs) (see the upper-right side of Fig. 3.1). In Scenario 1, the dataset was processed to eliminate the variables that were identified as the least important factors based on information gain, gain ratio, and chi-squared FSMs. The goal was to reduce the size of the dataset by eliminating the predictors that were within the bottom 2% of importance. In Scenario 2, the most important predictors were identified based on random forest FSM and only these predictors were used in the susceptibility analyses. The most important factors in this scenario were determined with a consistent approach as in Scenario 1, where approximately 2% was used as the threshold to separate the variables from least to more important. In Scenario 3, no FSM has been utilized to screen the dataset. Therefore, the analyses were performed with the entire dataset. Details on the FSM used in this study are provided in the next section.

Feature Selection Methods

Four feature selection methods have been used in this study to identify the most important variables to evaluate the landslide susceptibility in a given region, as discussed below.

Information Gain

Information Gain (IG) is an entropy-based feature evaluation method. It is defined as the amount of information provided by the feature items for the landslide susceptibility analysis. In other words, it measures the information obtained for category prediction by defining the presence or absence of a term (Forman, 2003). The formula for information gain is as follows:

Info Gain (Class, Attribute) = H (Class) – H (Class | Attribute) Equation. 3.1

Where H is the Entropy of the set and is defined as follows:

$$H(S) = Entropy\ (S) = -\sum_{j=1}^{j=m} P_j \log_2 P_j \qquad \text{Equation. 3.2}$$

Where P is the probability with which a particular value occurs in the sample space S. Entropy ranges from 0 (all instances of a variable have the same value) to 1 (equal number of instances of each value). Higher Entropy means the distribution is more uniform.

Gain Ratio

The Gain Ratio (GR) algorithm is a normalized version of the IG algorithm. The normalization is done by dividing the information gain by the entropy of the attribute with respect to the class, because it is claimed that this reduces the bias of the information gain algorithm. The formula for GR is as follows:

Gain Ratio (Class, Attribute) = (H (Class) – H (Class | Attribute)) / H (Attribute)

Equation. 3.3

Although GR is closely related to the IG algorithm, in this study a comparison was made to demonstrate the effect of normalization of gains in each attribute (or conditioning factor). The analyses were conducted to evaluate the effect of increased number of attributes in the subset to the difference in results from the GR and IG analyses.

Chi-Squared

Chi-Squared (CS) is another statistical method that is applied to test the independence of two events, where two events A and B are defined to be independent if P(AB) = P(A).P(B) or, equivalently, P(A|B) = P(A) or P(B|A) = P(B). X^2 in this method is calculated to show the difference between each feature and target. The higher X^2 values indicate features with higher importance. The formula on how the X^2 is calculated is as follows:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \text{Equation. 3.4}$$

where O_i is the observed frequency (number of observations) in class I, and E_i is the expected frequency if NO relationship existed between the variables.

Random Forest Importance

Random forest importance (RFI) is another feature evaluation method that can be used to rank the importance of variables. The RFI technique is described by Breiman L (2001), which involves first calculating the importance of each variable (predictor) and then normalizing these variables based on standard deviation of the differences in out-of-bag (OOB) error before and after the variables are permuted and compared within the training data. The OOB is a technique to measure the prediction error of RFI using the bootstrap aggregating (bagging) method in training samples and sub-samples of the dataset. During this process, error for each data point is recorded and averaged over the forest.

Susceptibility Algorithms

Three different algorithms have been used in this study. The input variables for these analyses were selected based on the feature selection results performed with four different methods as discussed previously. Analyses were compared both by selecting the top most important variables that were determined as a result of feature selection methods and by evaluating all of the variables (i.e., 21 predictors). Fig. 3.1 shows the application of these algorithms to evaluate the three different scenarios outlined in this study. Below provides brief description for each of these methods.

Random Forest

Random forest (RF) is one of the algorithms used to estimate landslide susceptibility analysis that can be applied to the classification and regression (Breiman, 2001). For each variable, the function determines model prediction error if the values of that variable are permuted across the out-of-bag observations (Trigila et al., 2015). In this study, the dependent variable has been represented by 0 and 1 for landslide and non-landslide pixels on the map, respectively. The RF feature selection method uses the bagging technique to select, at each node of the tree, random samples of variables and observations as the training dataset for model calibration. Since the random selection of the training dataset may affect the results of the model, using numerous trees help to balance the stability of the model. Unselected cases (out of the bag) are used to calculate the error of the model (OOBError), equal to the standard deviation error between predicted and observed values, and to establish a ranking of importance of the variables. It holds true that the greater the prediction error, the greater the importance of the variable.

Decision Tree C4.5 and C5.0

C4.5 is one of the most widely-used algorithms to classify the examples in a decision tree. This algorithm is designed to develop a tree for the binary dependent variable (landslide and non-landslide) (Quinlan, 1986). The decision tree is based on a multistage or hierarchical decision scheme (tree structure). The tree is composed of a root node, a set of internal nodes, and a set of terminal nodes (leaves). Each node of the decision-tree structure makes a binary decision that separates either one class or some of the classes from the remaining classes. The processing is carried out by moving down the tree until the terminal node is reached that only contains elements of a class. In a decision tree, features that carry maximum information are selected for classification, while remaining features are rejected, thereby increasing computational efficiency (Saito et al., 2009). The attribute selection measure uses the concept of entropy, which is defined as the degree of disorder. Thus, a tree grows by selecting an attribute with the smallest entropy or highest information gain (Yeon et al., 2010).

C5.0 is an updated version of C4.5. This model extends the C4.5 classification algorithms described in (Salzberg, 1994). Based on Kuhn and Johnson (2013), the updated version has been improved in terms of speed, memory usage, the size of the decision tree, boosting (improves the trees accuracy), and weighing which allows the user to weigh different cases and misclassification types. In a comparison study of the top 10 algorithms of data mining, (Wu et al., 2008) introduced C4.5 as the most influential data mining algorithm. In this study, both versions of this decision tree models were used and the results were compared.

Results and Discussion

Figure 3.2 presents the results of the application of the feature selection methods. The results from the IG, GR, and CS algorithms were similar. All these methods identified the slope percentage, the type of bedrock, and the percentage of sand in the first layer as the least important factors (see Figures 3.2a, 3.2b, 3.2c). A potential reason for eliminating slope percentage from the dataset combination based on FSM analyses could be the high percentage of clay in the landslide areas within this region. Based on results of Tommasi et al. (2012), in landslides with over-consolidated clays, slope percentage does not play an important role in triggering landslides. Figure 3.2d presents the evaluation of the factors based on the random forest features selection method. The least important factors identified with this method are different from those identified with the other three FSMs. Therefore, we considered two scenarios: Scenario 1 corresponding to the features selected by the IG, GR, and CS algorithms, and Scenario 2 corresponding to the features selected by the random forest algorithm (see the right-hand side of Fig. 3.1). In order to evaluate the impact of these FSMs to the output of the landslide susceptibility analyses (LSA), each scenario has been evaluated by itself (raster vs. vector data). The results were then compared with the existing landslide information that was available for the study area. This information was then used to quantify the accuracy and error associated with each LSA. Also, to determine the performance of each dataset, the outcomes obtained from the three different scenarios were compared to each other.

The results of the LSA were evaluated for the areas both within and outside of the regions where the landslides existed. The results from areas with landslides were presented as true positive ratio (TPR) and false negative ratio (FNR). The TPR values represent the correct prediction of the analyses and FNR values represent the incorrect prediction of the analyses. Hence, as the TPR values increased, the FNR values are expected to decrease for a more accurate result. The results from areas without landslides were also evaluated similarly, however; in that comparison the used terms were true negative ratio (TNR) and false positive ratio (FPR). Therefore, the increase in TNR value represents the correct interpretation of the non-landslide areas. As the TNR increased, FPR is expected to decrease. The approach on how the predictions have been validated is shown in Fig. 3. The next sections describe the results of these comparisons obtained from each of the three algorithms.

Prediction Analyses Results with Random Forest for All Scenarios

Table 3.1 presents the results of all the analyses conducted with the random forest algorithm. When the results from the data generated by raster and vector approaches were compared, the accuracy of the results varied within and outside of the landslide areas. In the landslide area, the results (as indicated by TPR) showed that the vector dataset led to better predictions on the existence of the landslides within the given area (i.e., high TPR and low FNR values). All the scenarios showed similar results although the results from scenarios 1 and 3 were slightly better than those from scenario 2. In the non-landslide area, the results from the raster dataset led to better predictions, as indicated by the very high TNR and very low FPR values. There were no significant differences between the predictions obtained from the three scenarios.

Prediction Analyses Results with C4.5 Algorithm for All Scenarios

Table 3.2 indicates the results for all analyses performed with the C4.5 LSA model. Comparing the results from both data sets generated by using raster and vector approaches showed that the percentage of correct predictions of landslides was higher with vector dataset. Also, among the three scenarios, the Scenario 3 had the highest TPR value for predicting landslide areas. In the non-landslide areas, the performance of raster data had a higher edge as compared to the vector dataset generated for this analysis. The results from Scenario 1 and Scenario 3 showed better results than the results from Scenario 2.

Prediction Analyses Results with C5.0 for All Scenarios

Table 3.3 presents the obtained results. When results from different datasets (vector and raster datasets) were compared, the order of the quality of predictions with C5.0 algorithm was similar to that obtained with C4.5algorithm. However, in terms of the TPR and TNR percentages, the results from C4.5 algorithm were better than those from C5.0algorithm.

Accuracy and Error Comparisons of Results from All Scenarios

Table 3.4 presents the overall accuracy and errors of all the scenarios evaluated based on both vector and raster datasets from all three LSA models and all three scenarios. The accuracy of the results presented in this table is defined by the ratio of correct predictions (i.e., combination of the number of true negative - TN and true positive – TP values) to the total number of predictions. In other words, it is determined from the ratio of the total number of correctly predicted landslide and non-landslide areas to the total number of observations. The misclassification error is determined by the ratio of the total number of false negative (FN) and false positive (FP) predictions to the total number of observations. It could also be computed by the complement of the estimated accuracy of the results. The values for TN, TP, FN, and FP were all presented in Tables 3.1, 3.2, and 3.3 for each scenario and each algorithm used in this study. Therefore, the data for Table 3.4 have been obtained from these tables. In this study, the total number of observations with the raster dataset was 3,645,234, and with vector dataset was 300. Therefore, there was a major difference between these two types of datasets. In the raster dataset, the number of data points that represented the non-landslide areas was 3,593,353 and the landslide areas was 51,881. The ratio between the non-landslide and landslide data points was approximately 69 to 1, which is considered as an unbalanced dataset. However, for the vector dataset, this ratio was 1 to 1, which is considered as a balanced dataset.

The results presented in Table 3.4 show the following:

- Even though the raster dataset is unbalanced, comparing the outcomes from all three LSA models, the results using the raster dataset had higher accuracy than those with vector datasets.
- When the outcomes from all three LSA predictions with raster dataset were compared, the accuracy of all analyses was close to each other, although the results

from C4.5 algorithm (especially for scenario 3) were better than all others. Also, for this particular case, the % error was the lowest.

- In all the LSA predictions with raster dataset, the results from scenario 3 were better than the results from the other two scenarios.

Conclusions

This study presented an experiment of performing landslide susceptibility analyses based on three different algorithms (RF, C4.5, and C5.0), two different data types (raster and vector), and three different data combinations (scenarios 1, 2, and 3) that were determined by using four different feature selection methods (IG, GR, CS, and RF). Identifying the locations of existing landslide areas by spending time in the field is a major task which is not always granted in large projects. The study made an attempt to identify a single algorithm as well as a dataset combination with the highest accuracy and prediction capacity. The following are the specific findings and limitations of the study:

- The results presented in Table 3.4 for vector dataset show that RF has lower accuracy than C4.5 and C5.0. These findings indicate that although RF is a standard and common algorithm to estimate landslide susceptibility analyses, the C4.5 and C5.0 algorithms have much more accurate outcomes.

- Even for the RF algorithm that has lower accuracy than C4,5 and C5.0algorithms, the results for raster data has a higher edge to the vector dataset. This indicates that raster data improves the results for the RF algorithm by taking thousands of pixels into account for the study area.

- Although using raster data has improved the results in RF analyses, this change of data type may not be as significant when it comes to C4.5 and C5.0. The main reason for a higher accuracy of the C4.5 and C5.0 algorithms might be the better performance for handling unbalanced data sets.

- Comparison of the results between different scenarios of raster data indicates that scenario 3 from all algorithm has a slightly higher accuracy. This finding shows that combining all parameters collected for the study has been helpful to increase the performance of the landslide susceptibility analyses.

A limitation of this study lies in the fact that the massive dataset of the susceptibility models requires high-performance computers to run the algorithms and compute the accuracy and misclassification errors which might be difficult in larger study areas. Thus, the size of the area to be evaluated has to be considered when selecting the computational process.

Finally, this study illustrates a methodology on how to compile the data for analysis, how to model the data within the analysis, and how to determine the most accurate algorithm for predicting the most probable landslide areas.

Tables

Table 3.1. Results of Analyses from Random Forest with All Scenarios

Random Forests Prediction Models	No. of Observations	TPR (%)	FNR (%)	TNR (%)	FPR (%)
		Landslide area		Non-landslide area	
Raster data					
Scenario 1	3,645,234	52.00	48.00	95.08	4.92
Scenario 2		73.80	26.20	96.76	3.24
Scenario 3		61.32	38.68	96.87	3.13
Vector data					
Scenario 1	300	84.00	16.00	88.67	11.33
Scenario 2		82.41	17.59	85.39	14.61
Scenario 3		84.67	15.33	90.00	10.00

Notes: TPR: True positive ratio (correct prediction of landslides),

FNR: False negative ratio (incorrect prediction of landslide),

TNR: True negative ratio (correct prediction of non-landslides), and

FPR: False positive ratio (incorrect prediction of non-landslides).

Table 3.2. Results of Analyses from C 4.5 with All Scenarios

C 4.5 Prediction Models	No. of Observations	TPR (%)	FNR (%)	TNR (%)	FPR (%)
		Landslide area		Non-landslide area	
Raster data					
Scenario 1	3,645,234	86.16	13.84	99.90	0.10
Scenario 2		83.12	16.88	99.90	0.10
Scenario 3		91.82	8.18	99.92	0.08
Vector data					
Scenario 1	300	96.67	3.33	96.67	5.33
Scenario 2		93.74	6.26	91.07	8.93
Scenario 3		97.33	2.67	94.67	5.33

Notes: TPR: True positive ratio (correct prediction of landslides),

FNR: False negative ratio (incorrect prediction of landslide),

TNR: True negative ratio (correct prediction of non-landslides), and

FPR: False positive ratio (incorrect prediction of non-landslides).

Table 3.3. Results of Analyses from C 5.0 with All Scenarios

C 5.0 Prediction Models	No. of Observations	TPR (%)	FNR (%)	TNR (%)	FPR (%)
		Landslide area		Non-landslide area	
Raster data					
Scenario 1	3,645,234	85.05	14.95	99.90	0.10
Scenario 2		81.22	18.78	99.73	0.27
Scenario 3		90.38	9.62	99.92	0.08
Vector data					
Scenario 1	300	93.33	6.67	91.33	8.67
Scenario 2		90.78	9.22	89.37	10.63
Scenario 3		94.67	5.33	92.00	8.00

Notes: TPR: True positive ratio (correct prediction of landslides),

FNR: False negative ratio (incorrect prediction of landslide),

TNR: True negative ratio (correct prediction of non-landslides), and

FPR: False positive ratio (incorrect prediction of non-landslides).

Table 3.4. Comparison of the Accuracy and Error Percentages from All Scenarios

Prediction Models	Scenario	Data Type	Accuracy %	Error %	Data Type	Accuracy %	Error %
RF	1	Raster	98.67	1.32	Vector	86.33	13.67
	2		97.32	2.68		86.33	13.67
	3		98.58	1.42		87.33	12.67
C4.5	1		99.72	0.28		95.67	4.33
	2		98.23	1.77		94.32	5.68
	3		99.82	0.18		96.00	4.00
C5.0	1		99.69	0.31		92.33	7.67
	2		98.48	1.52		91.31	8.69
	3		99.78	0.21		93.33	6.67

Figures

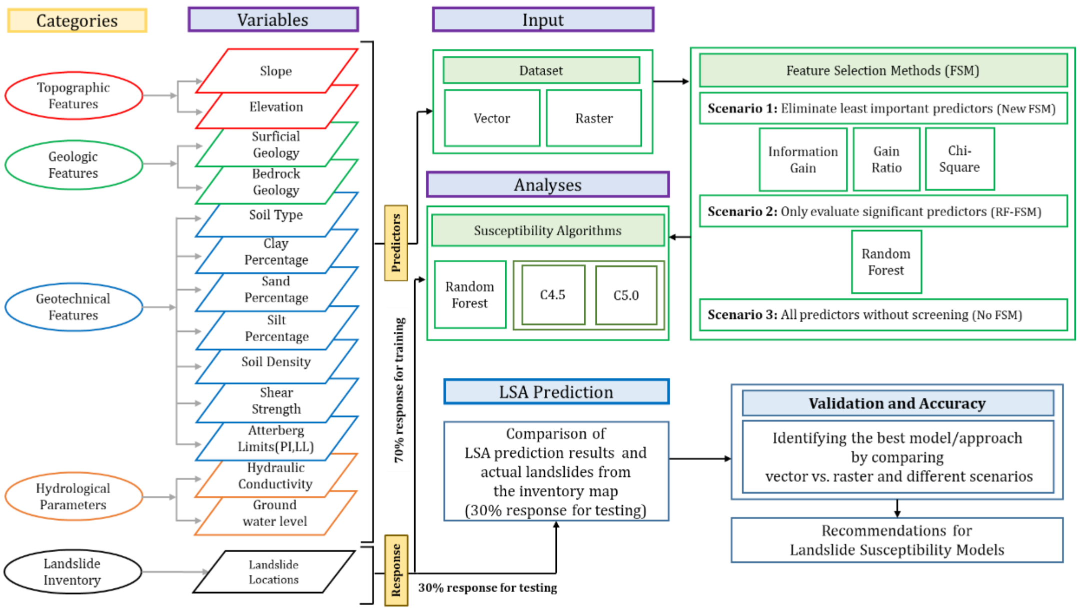


Fig. 1.1. Procedure of landslide susceptibility analysis using Feature Selection Methods and Decision Trees

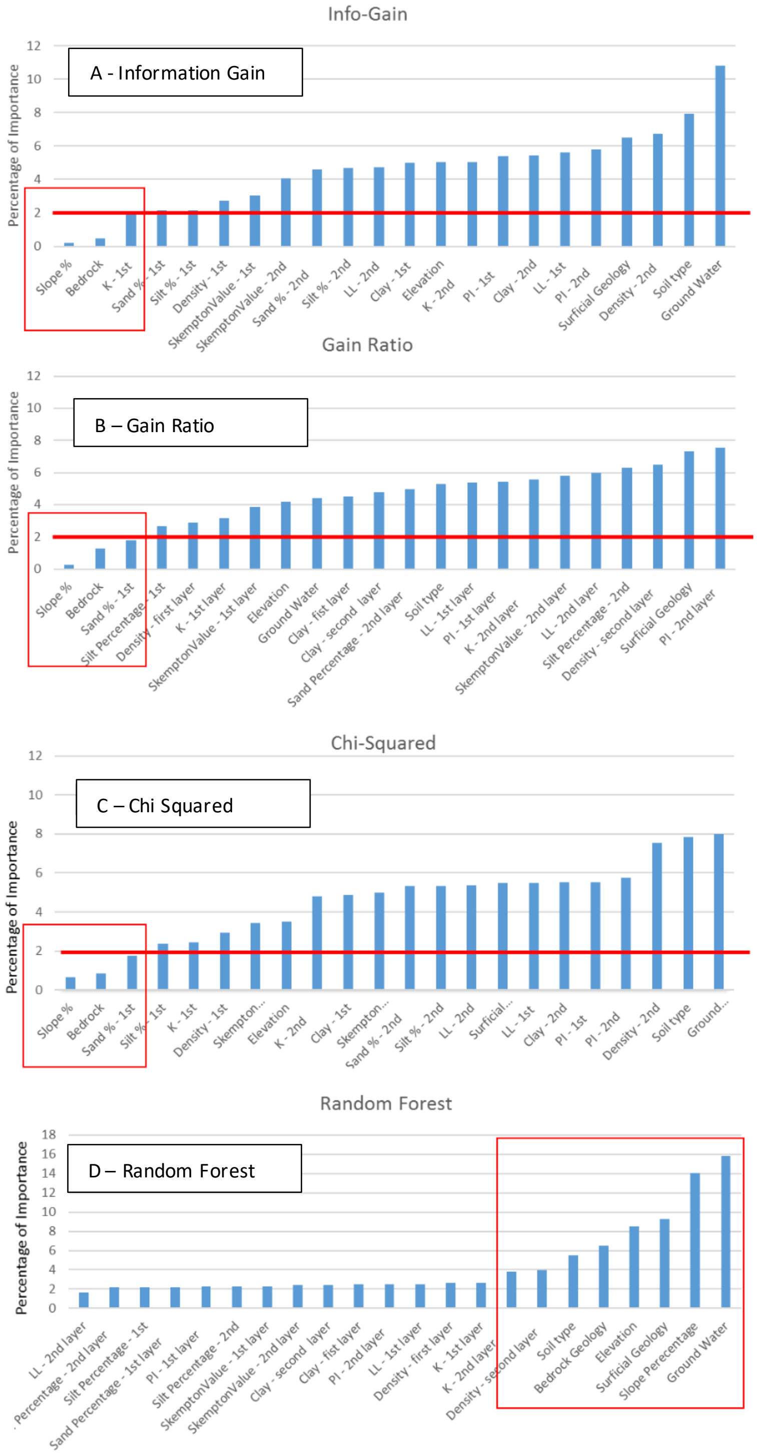


Fig. 3.2. Results of the Feature Selection Methods

CHAPTER 4: Use of C4.5 Machine Learning Algorithm to Predict Rainfall Induced Landslides[3]

Introduction

Hazard analyses consist of mapping and evaluating the potential for damage by incorporating external factors. Hazard analyses differ from susceptibility analyses by incorporating the factors that result in triggering landslides. Examples of previously conducted analyses with different approaches can be found in Dhakal and Sidle (2004), Baum and Godt (2010), and Rossi et al. (2013).

One of the basic techniques of landslide hazard analyses is to conduct an evaluation of the slope stability. This approach is heavily used by geotechnical engineers to estimate potential failure mechanisms of human-made or natural slopes (e.g., excavations, landfills, and roadways). Slope stability analysis may be classified into deterministic (factor of safety) and probabilistic approaches. The deterministic methods are mainly calculated based on a specific mathematical model and the physical properties of the slope. Most deterministic models utilize a limit equilibrium analysis for a defined failure surface (e.g., Swedish slip circle, modified Bishop, Spencer’s method, etc.). These models are commonly used in geotechnical investigation as mono- bi- and tri-dimensional approaches. The calculation of the factor of safety requires precise data on slope geometry, shear

[3] The research described in this chapter has been submitted as a journal manuscript in March 2020.

strength properties of the soil, and pore water pressure. Therefore, one of the main disadvantages of this approach is that the slope stability is evaluated by determining a worst-case scenario where slope and the associated parameters and features are drawn as part of a specific cross section in great detail. Additionally, the factor of safety is computed individually for each cross section of the slope. Thus, predicting shallow landslides over large regions in real or near real time is an impossible task based on the deterministic approach (Baum & Godt, 2010). The probabilistic methods include the analyses to be conducted with variety of geotechnical properties (i.e. cohesion, angle of internal friction, undrained shear strength) but is still based on a given cross sectional area (not covering large areas/regions). Therefore, whether the slope stability evaluation is conducted based on deterministic or probabilistic approaches, in both methods, the extent of the area included in the analyses is very limited and not suitable for a regional evaluation.

Physically based landslide analyses are more suited to evaluate larger areas. The appearance of geographical information system (GIS) and high computational ability of computers ease applying algorithms to every cell of the grid-based dataset to analyze distributed slope stability model. Several physically based approaches have been utilized previously to address landslide hazard analysis. SHALSTAB (Shallow Landslide Stability Model) is one of the grid-based approaches, which involve a hydrological flux coupled with an infinite slope analysis. However, this model is not appropriate to forecast the timing of slope failure based on the triggering mechanism (Dhakal and Sidle, 2004). SINMAP (Stability Index Mapping) is a different physically-based method that works based on infinite slope stability model with groundwater pore pressures. In this method, the required data for pore pressure are obtained from a steady state model. SINMAP allows an uncertainty of the variables through the specification of lower and upper bounds of each conditioning factors (Rossi et al., 2013). This method applies only to shallow transitional landslide phenomena controlled by shallow groundwater flow. It does not apply to deep-seated or rotational slides. Another disadvantage of this method is that the accuracy of output is heavily reliant on the accuracy of DEM (Yatheendradas et al., 2019). TRIGRS (Transient Rainfall Infiltration and Grid based Regional Slope stability) is another approach developed in FORTRAN language, for computing the transient pore pressure distribution due to rainfall infiltration using the method proposed by (Iverson, 2000) model (Baum and Godt, 2010). The results are stored in a distributed map of the factor of safety. However, TRIGRS is very sensitive to initial conditions, therefore, if the initial water table depth is poorly constrained, it may produce questionable results. All these approaches have a value to estimate landslide hazard analyses but they do not consider geotechnical/geological features within the landslide area (e.g., types of bedrock and soil, depths of stratigraphic layers, percentage of soils, and shear strength parameters of soil). Therefore, the analyses are predominantly solved by using rainfall data and topographical information of the slope which significantly limits the accuracy of the predictions.

The importance of rainfall to trigger landslides is a known fact (Iverson, 2000). However, if the rainfall intensity that might have caused a specific landslide in a region is known, this information can then be used to estimate the possibility of landslide occurrence in that region in the future. The purpose of this study was to determine such relationship between the rainfall intensity and the occurrence of landslides. To the best of the authors' knowledge no previous study has used such approach that is defined in this study to achieve this goal. In this study a region with known landslide locations and extensive dataset of soil properties associated with and without landslides were used to first train the machine learning algorithm and then to re-perform the analyses based on the new datasets that were created based on the rainfall infiltration effect. This approach allowed the estimation of the rainfall intensity that might have most likely caused the landslides that were observed in a specific region. The approach followed in this study considered geological, geotechnical, and hydrological properties, which in itself makes this study unique.

Model Development

Fig. 4.1 describes the approach used in this study to relate the intensity of the precipitation into surface water infiltration and triggering of the landslides that have previously been identified (or mapped). In order to conduct this study, following assumptions had to be made:

- All precipitation infiltrates into the soil layers (no runoff);
- Only two layers of soil exists in the region;
- Shear strength and density of the regions with no landslides initially had similar properties in terms of shear strength and density.
- Shear strength and density of the soil layers change as each layer gets inundated with water based on the estimated depth of infiltration; and

Properties of the study area

The analyses were performed by first developing a digital elevation model (DEM) of a region. This DEM was then used as a scale raster data set to embed all available properties within the region including areas with and without landslides. Raster data is a form of data set that contains pixels (cells).

The soil properties that are used in this study are shown in Fig. 4.1 as slope, elevation, surficial geology, bedrock geology, soil type, Atterberg limits (this information states about the consistency of the clayey soil), and different percentages of sand, silt, and clay sized particles within each soil layer. The study area consists of 15 different types of surficial geological features, ranging from what is referred as Wisconsin, glacial, and re-sedimented tills to alluvium that contained stratified silt, sand, gravel, and boulder size particles. There were three different types of bedrock identified at the site consisting of Huntley Mountain Formation, Catskill Formation, and Lock Haven Formation. All of this data has been obtained from PADCNR (Delano and Wilshusen, 1999; Pennsylvania Department of Conservation and Natural Resources, 2019) and USDA (United State Department of Agriculture, 2019) in different depths, which allowed the model to be formulated to have two different layers. The first soil layer has a thickness of 30 cm (12-inches) and the thickness of the second layer is assumed to extent all the way to the top of bedrock. The definition of these thicknesses was defined based on the available data and is supported by the fact that most of the landslides in this region were previously identified to be surficial, where the surface conditions of the slope impact the stability (Delano and Wilshusen, 1999). Although the properties of the soil were different with depth, they also varied spatially. However, once they were assigned to a pixel, these properties stayed the same throughout the analyses. Additionally, all pixels that were located within non-landslide areas were represented by an effective peak friction angle and pixels within landslide areas were represented by residual friction angles. This approach is consistent with what is typically observed by geotechnical engineers in their probabilistic analyses (Duncan 2014). All layers without groundwater were represented by moist density (Tarboton, 2003). Table 4.1 shows the range of the friction angle and moist density values that were obtained from the literature based on the soil type at a given location (Loehr et al., 2017).

The hydrological properties of the study area are also shown in Fig. 4.1 as groundwater level, hydraulic conductivity, soil porosity, effective porosity, wetting front soil suction head, and initial effective saturation. The groundwater level in the study area was mapped out based on the static water levels (meter below the surface) provided by the Pennsylvania Spatial Data Access (PASDA) from each of the wells located within the area. The static water level was added to the ArcGIS as a vector layer with the location and depth of each well. This information was then used to create a contour map and raster layer of the groundwater level. The rest of the hydrological properties have been obtained from the literature based on the type of soil defined at each location. Table 4.2 presents the range of hydrological properties data based on the soil type at a given location (Tarboton, 2003).

Rainfall scenarios considered within the study area

Rainfall and precipitation are the major triggering factors in this study area (Delano and Wilshusen 1999). To study the effect of different historic rainfall on the hazard analyses, the precipitation data were collected from Tioga Hammond Dam data accumulation center from 2000 to the late 2018 (United States Geological Survey, 2019). This data center is in approximately 10 kilometers of the Mansfield city. Fig 4.2 shows the precipitation data plotted based on the accumulative rate centimeter per day. For the purpose of this study, three different scenarios were selected to represent the intensity of the rainfall in different ranges: 1) intense rainfall, 2) medium-range rainfall, and 3) No rainfall. The rainfall data for intense, mid-range and no rainfall days were collected on 9/8/2011 at 104.6 mm, 5/22/2018 at 23.9 mm, and 12/20/2006 at 0 mm respectively. These days have been selected randomly from the data acquired from Tioga Hammond Dam data accumulation center (Fig. 4.2). By selecting these three rainfall scenarios, the goal of the study was to estimate their infiltration rate into the ground and evaluate the hazard analyses in three different rainfall conditions.

Methodology Used to Estimate Depth of Rainfall Infiltration

Depth of rainfall infiltration has been estimated based on the Green-Ampt (GA) infiltration model. The GA model is a simplified representation of the infiltration process, which solves the Richard's equation used to estimate the water infiltration rate into the soil. This model assumes a homogenous soil profile and a uniform distribution of initial soil water content. Based on different soil types, GA model uses soil porosity (ε), effective porosity (θ_e), wetting front soil suction head (ψ), initial effective saturation (σ_e), and hydraulic conductivity (K) to calculate the infiltration rate (f). Hence these properties had to be pre-defined for the region where the landslides and no landslides exist. Table 4.2 based on previous literature is created for this purpose (Tarboton, 2003).

The infiltration rate (*f*) (cm/hr) is estimated by the GA model based on the equation 4.1 as provided below:

$$f = K \left[\frac{\psi \Delta\theta}{F_p} + t\right] \qquad \text{Equation 4.1}$$

where equation 4.1 is a function of equations 4.2 through 4.4:

$$\Delta\theta = \theta_e (1 - \sigma_e) \qquad \text{Equation 4.2}$$

$$F_p = i t_p \qquad \text{Equation 4.3}$$

$$t_p = \frac{K \Psi \Delta\theta}{i(i-K)} \qquad \text{Equation 4.4}$$

$\Delta\theta$: difference between initial and the moisture content equivalent to effective porosity

F_p: depth of cumulative infiltration from the ponded water (cm)

i: rainfall intensity that is of interest based on different scenarios (cm/hr)

t_p: time of ponding of the water at the surface (hr)

t: duration of interest for the infiltration rate (hr)

Once the infiltration is known then the infiltration depth *F(t)* is calculated as shown in equation 4.5 as a function of *t*:

$$F_t = f \times t\ (hr) \qquad \text{Equation 4.5}$$

In the overall model developed for this study, at each pixel within each soil layer, the equation for the GA model has been coded in order to estimate the depth of infiltration (Eq. 4.5). To simulate the effects of the different rain intensity, the corresponding rainfall data is also model at each pixel level. Based on this information, the GA model estimated the depth of infiltration. After this information is obtained, again at each pixel level, the estimated depth of infiltration was compared against the thickness of soil layers. If the depth of the rainfall infiltration was estimated to be greater than the thickness of the first layer then the friction angle was revised from being effective to residual and the density from moist to saturated. In this process, the existence of the pre-determined landslide locations have been disregarded in regards to the friction angle and density. This approach was necessary to create a new dataset with newly assigned friction angle and density values but by keeping all other pre-assigned soil properties constant (see Fig. 4.1 for the logic order). This approach was necessary because the purpose of this study was not to predict the existing locations of the landslides but to predict the potential to trigger new landslides even in areas where in the past landslides have occurred.

Decision Tree Model to Predict New Landslides

The field study conducted by PADCNR shows the locations of existing landslides. However, the rainfall conditions that might have triggered these landslides are not known. The goal of this study was to determine the most likely rainfall scenario that can trigger these landslides. It is almost impossible to predict the future, however, if the rainfall intensity that might have triggered the existing landslides in a given region is known, then this information may lead to estimate future occurrence of landslides. This is based on the assumption that the rainfall intensity that caused landslides in the past may also cause landslides in the future. To perform such analyses, a machine learning algorithm is needed. In this study, such analyses have been conducted with C4.5 algorithm (Quinlan, 1993). This algorithm was selected because the previous study conducted by Alimohammadlou et. al (2020) has showed that C4.5 algorithm had the best accuracy to be used in studies that required the use of machine learning approaches in landslide analyses.

The working principles of C4.5 algorithm is based on developing a tree for the binary dependent variables (landslide and non-landslide). The decision tree is based on a multistage or hierarchical decision scheme that is based on nodes. These nodes composed of a root node, a set of internal nodes, and a set of terminal nodes (leaves). Each node of the decision-tree structure makes a binary decision that separates either one class or some of the classes from the remaining classes. The processing is carried out by moving down the tree until the terminal nodes are reached. In a decision tree, features that carry maximum information are selected for classification, while remaining features are rejected, thereby increasing computational efficiency (Saito et al., 2009). Thus, a tree grows by selecting an attribute with the smallest entropy or highest information gain (Yeon et al., 2010).

In this study, a model has been created with C4.5 algorithm that included all of the dataset with soil properties of the regions with and without landslides. This model was used to train C4.5 algorithm to differentiate between the properties that are observed in areas with and without landslides. This trained model was then used to predict the occurrence of new landslides based on the dataset that was created as a result of the GA infiltration analyses (see Fig.4.1 for the logic order). If the results of these analyses showed no new landslides, this meant that the selected rainfall intensity was not high enough to infiltrate into the soil layers or changing the soil properties. In the case where the rainfall intensity resulted in infiltration into the soil layers and changing the soil properties, the results showed new landslides. In the case where the new landslides occurred at the same locations that were previously defined by PA DCNR, the results were interpreted as the rainfall intensity that was simulated in this study to be similar to the rainfall events that might have happened in the past that have resulted in forming the landslides

Results and Discussion

Fig. 4.3 shows the outcome of the GA model in all three regions in this study. The figure shows the areas within each region where surface infiltration has affected the soil properties. As can be seen from the figure, there was no effect on soil properties when there was no rainfall (as expected). However; with the increase of rainfall intensity, the surface water has penetrated further into the soil layers and traveled deeper into the ground. Fig. 4.3a, 4.3b, and 4.3c show the extent of the infiltration after medium and high intensity rainfall events in three different regions respectively. As can be seen from Fig. 4.3a for the abundant landslide area, in the case of the medium intensity rainfall only 22.18% of the area had rainfall infiltration into the first layer (top surface layer) and 10.64% into the second layer. During the high intensity rainfall, these percentages increase to 57.25% and 28.76% respectively. These changed conditions, create a new dataset for the regions where the soil friction angles and the density are changed due to the fact that the layers got inundated with infiltrated water.

Fig. 4.4 shows the newly predicted landslides in all three regions after the new dataset that is developed based on infiltration model (and the changed soil density and friction angle) are compared to the original dataset that was used to train the C4.5 model. Keeping mind that the analyses were performed on a pixel level, each white pixel (or cell) indicates a location where a new landslide is predicted. The triggering factor in these analyses is the infiltration of water into the soil layers as this resulted in changing soil properties. However, the actual determination of where a new landslide may or may not occur is much more complex. This is because C4.5 model incorporates all landside features that it has been trained with to determine the prediction of the new landslide locations. Therefore, such analyses require a similar model that is developed in this study.

The results of the C4.5 model based on no rainfall scenario is presented in Fig. 4.4a and summarized in Table 4.3. Because there is no infiltration in this rainfall scenario, no changes have occurred in the soil properties, therefore it is expected that no newly developed landslide areas should be predicted. When the results were compared from the three different regions (as defined by three different tiles), in all regions some pixels show new predictions. However, the percentage of these newly predicted landslides are 0.61, 0.89, and 2.26% of the overall study area in no-landslide, medium-range, and abundant landslide areas respectively. These predictions are considered as the noise (i.e., the misclassification of the C4.5 algorithm), resulting in false positives. Therefore, when interpreting the overall results from this study, it should be kept in mind that the accuracy of the predictions may possibly range ±2.5%. Based on the previous study conducted by Alimohammadlou et. al (2020), the overall accuracy of the C4.5 model for landslide analyses was estimated as 99.82% when the model was used to estimate the susceptibility of existing landslides. Therefore, it is not of surprise to observe some false positives that occur in analyses conducted to predict new landslides.

The effect of the rainfall to trigger new landslides become evident even in the medium rain intensity scenario (Fig. 4.4b) as can be seen by the coverage of the white pixels in the areas. In the case of abundant landslide areas, the model predicts the triggering of some of the existing areas that already have landslides as well as a small percentage of new areas to also trigger landslides. The newly triggered landslides cover 4.14 percent of the total area of the tile. However, in the case of medium-range landslide areas, the percent of newly triggered landslides is much less and only appearing to cover the area with previously existing landslide regions. This area covers 1.92 percent of the area of tile number 61002140. In the area where no previous landslides exist, the model still predicts that in some areas new landslides will trigger. The percentage of these newly triggered areas cover 1.33 percent of the total area of the tile number 61002160. Considering that the possible accuracy of the model is ±2.5%, predicted percentages less than 2.5% may also be considered as false positives but at this stage without being able to compare the predictions with an actual event (where medium intensity rainfall hits the area of medium range landslide area), it is impossible to differentiate the magnitude of this difference. However, the model indicates that there could be a potential risk for the triggering of landslide even in areas that previous landslides have not been reported.

The extent of the newly predicted landslide areas with high intensity rainfall scenario is shown in Fig. 4.4c. When compared with results from the medium intensity rainfall scenario two observations are noted: (1) the coverage of the newly predicted landslides increases dramatically and (2) the definition of the newly predicted landslide areas become more evident (the white pixels in each location become larger and better defined). Overall, as in the case with medium intensity rainfall scenario, new landslides have been predicted even in the area where previously no landslides were reported (Table 4.3). The predicted percentages in high intensity rainfall event at all areas were larger than the false positive threshold (i.e., 2.5%) indicating that a potential risk of occurring new landslides increases dramatically with the increase in rainfall intensity.

Conclusions

This study presented a methodology where a machine learning algorithm can be used to predict the possible future occurrence of landslides in a given region that is triggered by the rain events. The results presented in this study were not meant to be an absolute outcome as the predictions cannot be verified unless an actual new landslide occurs in the future within the specified study area. However, the study presented herein demonstrates that such methodology can be used to make predictions and also assess the potential bias within the predictions (i.e., predicting landslides in no rain events). Although not demonstrated herein, these predictions can then be used as early warning systems for the residents and municipalities that are located in landslide prone regions. This study specifically differs from the previous studies in many ways, where in this study, the predictions are based on the combination of actual geological, geotechnical, and hydrological properties of the region. Previous studies primarily focus on changes of the geometrical features based on rainfall events. Geotechnical engineers are able to conduct specific evaluations of a given slope but this study expands this capability at a regional level (covering very large areas such as an entire County within a region). The following are the specific findings and limitations of the presented study:

- The intensity of the rainfall plays a major role in terms of the depth of infiltration of the surface water into the geological formations.
- The geotechnical properties of the ground are complex and the properties that represent each geological layer must be selected by the trained person who understands the meaning of these properties.
- The infiltration of the rainfall may be predicted by an existing model, however hydraulic properties of the geological layers must be carefully evaluated/selected.
- Predicting of landslides require the machine learning algorithm to be trained based on the actual landslide data set prior to be relied on.
- C4.5 algorithm has been used in this study based on the previous findings of another study that has compared the accuracy of different decision tree algorithms. C4.5 algorithm was pointed out as the most accurate of previously tested algorithms for landslide studies.
- The methodology presented herein was kept simplistic by design to make the analyses less sophisticated (easier to implement), however, the results obtained with these simplifications might affect the accuracy. For example, the infiltration model used in this study was a simplified version of the Richard’s equation (Ross, 1990) that has been generally used to estimate the water infiltration in semi-saturated media.

- Although two rainfall intensity scenarios were evaluated in this study, the methodology presented would allow the user to input any rainfall intensity event in the analyses.

Tables

Table 4.1. Estimation of residual friction angle for clayey soils in the study area

Soils with clay fraction equal or less than 20 %

LL	Residual Phi Prime
80	17.5
60	22
40	26.5
20	32

Soils with clay fraction between 20 % and 45%

LL	Residual Phi Prime
140	9
120	10
100	12
80	16
60	19
40	24

Soils with Clay Fraction equal or greater than 45%

LL	Residual Phi Prime
290	3.5
280	4
260	5.5
240	6
220	6
200	6.5
180	7.4
160	8
140	9
120	9.5
100	10
80	12
60	17
40	20

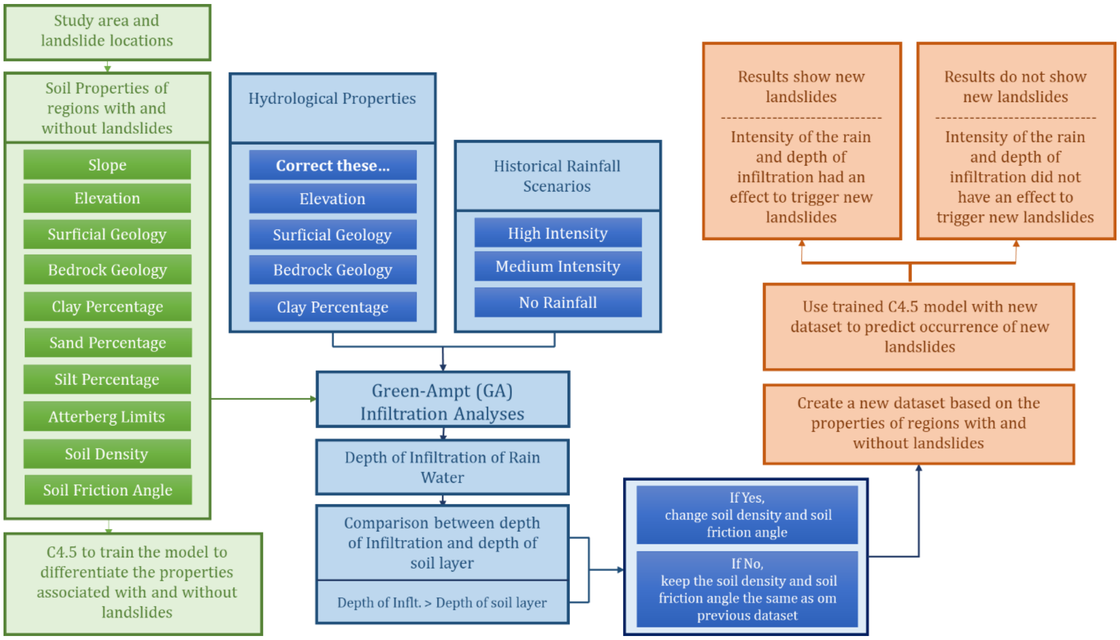
Table 4.2. Green – Ampt infiltration parameters for various soil classes.

soil texture	Porosity (n)	Effective porosity (θe)	Wetting front soil suction head (ψ) (cm)	Hydraulic conductivity (K) (cm/hr)
Sand	0.437	0.417	4.95	11.78
Loamy sand	0.437	0.401	6.13	2.99
Sandy loam	0.453	0.412	11.01	1.09
Loam	0.463	0.434	8.89	0.34
Silt loam	0.501	0.486	16.68	0.65
Sandy clay	0.398	0.33	21.85	0.15
Clay loam	0.464	0.309	20.88	0.10
Silty clay loam	0.471	0.432	27.30	0.10
Sandy clay	0.43	0.321	23.90	0.06
Silty clay	0.479	0.423	29.22	0.05
Clay	0.475	0.385	31.63	0.03

Table 4.3. Percentage of areas covered by new landslides as predicted based on C4.5 model as a function of three different rainfall scenarios.

	High Intensity Rainfall	Medium Intensity Rainfall	No Rainfall
Abundant Landslide	10.78%	4.14%	2.16%
Mid-Range Landslide	6.18%	1.92%	0.89%
No Landslide	3.06%	1.33%	0.61%

Figures



Study area and landslide locations
Soil Properties of regions with and without landslides
Slope
Elevation
Surficial Geology
Bedrock Geology
Clay Percentage
Sand Percentage
Silt Percentage
Atterberg Limits
Soil Density
Soil Friction Angle
C4.5 to train the model to differentiate the properties associated with and without landslides
Hydrological Properties
Correct these...
Elevation
Surficial Geology
Bedrock Geology
Clay Percentage
Historical Rainfall Scenarios
High Intensity
Medium Intensity
No Rainfall
Green-Ampt (GA) Infiltration Analyses
Depth of Infiltration of Rain Water
Comparison between depth of Infiltration and depth of soil layer
Depth of Inflt. > Depth of soil layer
If Yes, change soil density and soil friction angle
If No, keep the soil density and soil friction angle the same as om previous dataset
Create a new dataset based on the properties of regions with and without landslides
Use trained C4.5 model with new dataset to predict occurrence of new landslides
Results show new landslides
Intensity of the rain and depth of infiltration had an effect to trigger new landslides
Results do not show new landslides
Intensity of the rain and depth of infiltration did not have an effect to trigger new landslides

Fig. 4.1. Procedure of landslide hazard analysis using Green-Ampt infiltration model and Decision Tree C4.5

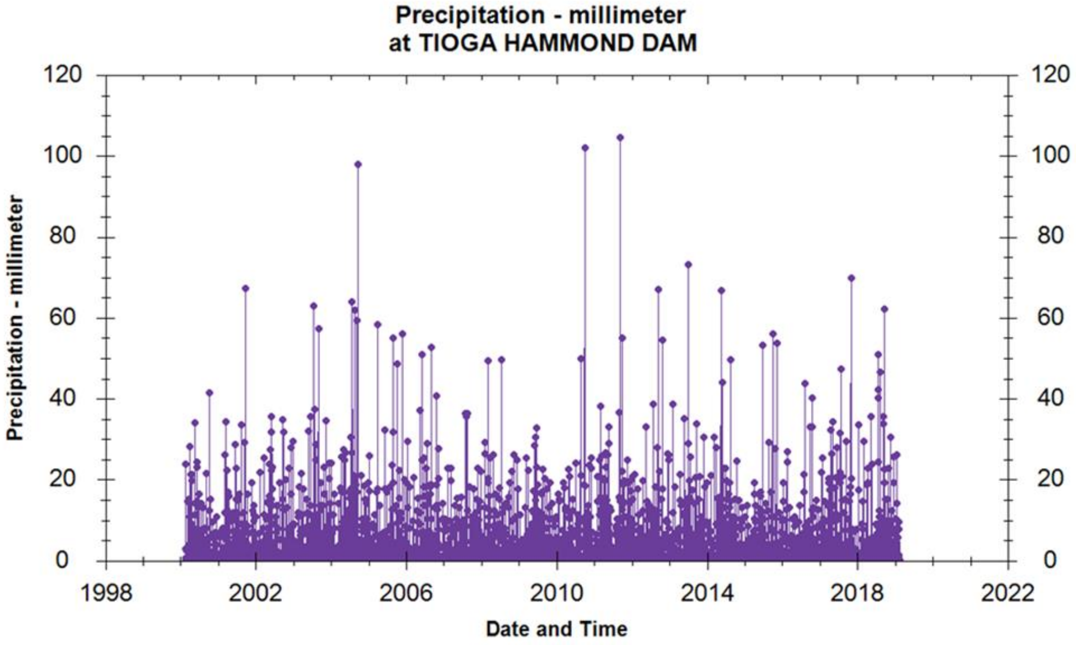
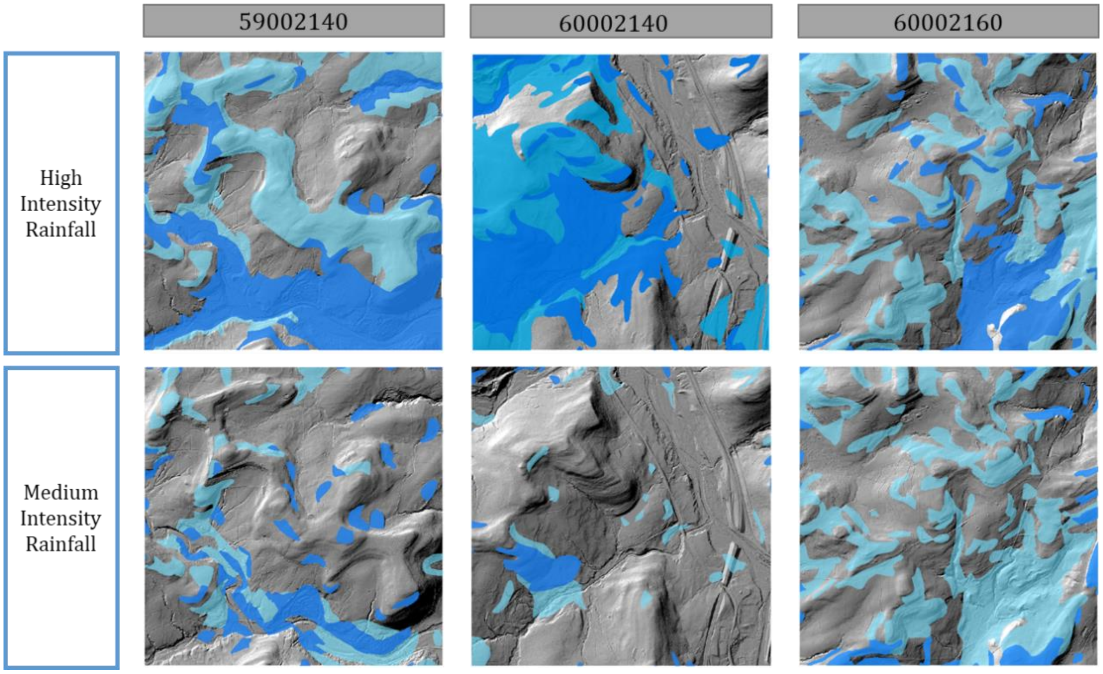


Fig. 4.2. Rainfall Data – Collected from Tioga Hommand Dam Data Center



(a) (b) (c)

Notes: Dark blue color shows the rainfall infiltration into the first (surficial) layer and light blue color represents rainfall infiltration into the second ground

Fig. 4.3. Rainfall infiltration rate as interpreted by the outcome of the Green-Ampt model in all three regions in this study area based on two different rainfall intensities

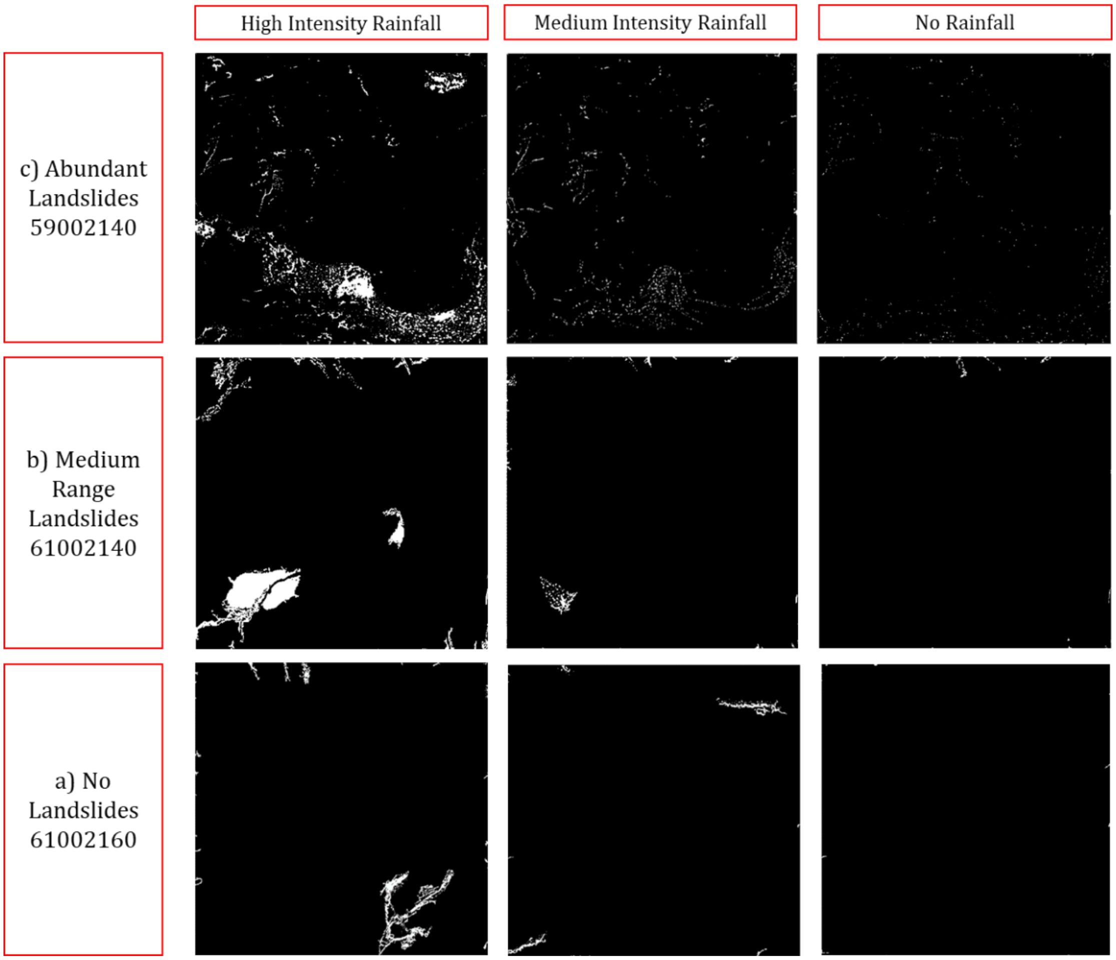

High Intensity Rainfall
Medium Intensity Rainfall
No Rainfall
c) Abundant Landslides 59002140
b) Medium Range Landslides 61002140
a) No Landslides 61002160

Fig. 4.4. Newly predicted landslides in all three regions within a) abundant (Tile No. 59002140), b) medium range (Tile No. 61002140), and c) no (Tile No. 61002160) landslide areas

CHAPTER 5: FINAL REMARKS AND CONCLUSIONS

Based on the findings and results discussed in the previous chapters, the initial hypotheses which this research was based on might be addressed with following conclusions:

The results of the landslide inventory analyses for the selected region reveals that the automated landslide detection model (ALDM) was able to capture both the landslides and non-landslides areas with accuracies of 70% and 92% respectively. This allows the evaluation of large areas with minimal effort in very short time as compared to visual detection methods. The results obtained were compared against data from PADCNR and landslides that were determined visually from the Hill shade map (a technique that is implemented by some of the DCNR agencies to delineate landslides). These findings support the first hypothesis. The study also reveals that the proposed ALDM method could be implemented in regions where the landslide sizes and features could be significantly different. This method requires the user to initially set thresholds for roughness and texture segmentation analyses, however once these thresholds are established, the analyses are conducted in large areas without supervision and the need to re-define these thresholds. On the other hand, although the ALDM was effective in capturing the younger landslides (i.e., less than 25 years), the method is not as effective in capturing the older landslides (landslides that have occurred 500 years or older). This is because although the thresholds could be redefined, the texture segmentation analysis has a limitation. Considering that in most cases the real dangers of landslides are associated with younger landslides that can remobilize, the proposed method is a viable technique to delineate landslides.

The results of landslide susceptibility analyses for comparing the raster versus vector data set reveals that predictions made with raster datasets provide a much more accurate outcome, even though raster datasets are considered as imbalanced data set. Therefore, the accuracies of all LSA results were high and the errors were low with raster dataset and this shows the importance of using raster datasets for landslide analyses. When compared, being able to implement all of the available variables (without eliminating any variable) into creating a dataset also improves the accuracy of the model. In all cases, the results obtained from scenario 3 (all landslide parameters included) had demonstrated this importance. This study aimed to evaluate the performance of decision tree algorithms (C4.5 and C5.0) in imbalanced data sets and compare the results to Random Forest model which is a standard and commonly used method in landslide susceptibility analyses. Comparison of the results from raster datasets with scenario 3 from all algorithms showed that the C4.5 algorithm had a higher percentage of being able to correctly predict both landslide and non-landslide areas. Therefore, the second and third hypotheses have been supported by these findings. It should be noted that the massive dataset of the susceptibility models requires high-performance computers to run the algorithms and compute the accuracy and misclassification errors which might be difficult in larger study areas. Thus, the size of the area that will be evaluated has to be considered when selecting the computational process.

The results presented in the hazard analyses demonstrate that methodology can be used to make predictions and also assess the potential bias within the predictions (i.e., predicting landslides in no rain events). This study specifically differs from the previous studies in many ways, where in this study; the predictions are based on the combination of actual geological, geotechnical, and hydrological properties of the region. Previous studies primarily focus on changes of the geometrical features based on rainfall events. Geotechnical engineers are able to conduct specific evaluations of a given slope but this study expands this capability at a regional level (covering very large areas such as an entire County within a region). Although this study were not meant to be an absolute outcome as the predictions cannot be verified unless an actual new landslide occurs, these predictions can then be used as early warning systems for the residents and municipalities that are located in landslide prone regions. Based on the fact that this study presented a methodology to predict possible future occurrence of landslides in a given region by using machine learning algorithm, the fourth hypothesis has been supported.

Limitations of this study and suggestions for future studies

The present study has been conducted in a study area located in Mansfield region of Pennsylvania. The goal was to develop a framework to quickly perform landslide inventory, susceptibility, and hazard analyses using machine learning algorithms and based on obtained DEM and landslide data. The findings of this study are inevitably limited to the accuracy of the data obtained from multiple resources, despite the fact that every effort was made to perform an accurate data pre-processing step. An attempt was made to validate the results of each section against the ground truth data obtained from Pennsylvania Department of Conservation and Natural Resources. Additionally, the results from inventory analyses were compared against the visual detected method. These comparisons were limited and might be a source of limitation for this study.

The hazard analyses were conducted by using three historical rainfall scenarios to predict potential future landslide in the given region. Real time rainfall scenarios are suggested for future researches to confirm the accuracy of landslide predictions. Although this study was conducted based on the soil data in two different depths, for the future studies, one suggestion is to increase the information within the database by adding soil data from additional depths. This could very well be achieved in the future for example, if any given particular county with the landslide active regions in the U.S. starts to create a repository of all of their boring logs used in their site investigations. Perhaps such efforst could be initiated as the study herein outlines a methodology on how to utilize such data for forecasting landslides. However, such effort would require machine learning algorithms perhaps using millions of data on supercomputers.

REFERENCES

Alimohammadlou, Y., Najafi, A., & Yalcin, A. (2013). Landslide process and impacts: A proposed classification method. *CATENA*, *104*(Supplement C), 219–232. https://doi.org/10.1016/j.catena.2012.11.013

Alimohammadlou, Y., Tanyu, B.F., Tecuci, G., 2019. Landslide Susceptibility Analyses Based on Random Forest, C4.5, and C5.0 Algorithms Using Balanced and Unbalanced Datasets. CATENA. Under Review.

Alimohammadlou, Y., Tanyu, B.F., Leshchinsky, B.A., 2019. A Novel Detection Model for Identification of Existing Landslides. Computers and Geoscience, Under Review.

Barlow, J., Martin, Y., & Franklin, S. E. (2003). Detecting translational landslide scars using segmentation of Landsat ETM+ and DEM data in the northern Cascade Mountains, British Columbia. *Canadian Journal of Remote Sensing*, *29*(4), 510–517. https://doi.org/10.5589/m03-018

Baum, R. L., & Godt, J. W. (2010). Early warning of rainfall-induced shallow landslides and debris flows in the USA. *Landslides*, *7*(3), 259–272. https://doi.org/10.1007/s10346-009-0177-0

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of Geographical Information Systems* (Third Edition). Oxford University Press.

Bolstad, P., 2012. GIS fundamentals: a first text on geographic information systems. White Bear Lake, Minn, 4th ed.

Booth, A. M., Roering, J. J., & Perron, J. T., 2009. Automated landslide mapping using spectral analysis and high-resolution topographic data: Puget Sound lowlands, Washington, and Portland Hills, Oregon. Geomorphology, 109(3-4), 132-147.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA*, *151*, 147–160. https://doi.org/10.1016/j.catena.2016.11.032

Dhakal, A. S., & Sidle, R. C. (2004). Distributed simulations of landslides for different rainfall conditions. *Hydrological Processes*, *18*(4), 757–776. https://doi.org/10.1002/hyp.1365

Duncan, J. M., Wright, S. G., Brandon, T. L. (2014). Soil Strength and Slope Stability. United States: Wiley.

Delano, H. L., Wilshusen J.P., 1999. Landslide susceptibility in the Williamsport 1- by 2-degree quadrangle, Pennsylvania.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*(Mar), 1289–1305.

Glenn, N. F., Streutker, D. R., Chadwick, D. J., Thackray, G. D., & Dorsch, S. J. (2006). Analysis of LiDAR-derived topographic information for characterizing and differentiating landslide morphology and activity. *Geomorphology*, *73*(1–2), 131–148. https://doi.org/10.1016/j.geomorph.2005.07.006

Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology, 31*(1), 181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., & Chang, K.-T. (2012). Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, *112*(1–2), 42–66. https://doi.org/10.1016/j.earscirev.2012.02.001

Highland, L. M., Godt, J. W., Howell, D. G., & Savage, W. Z. (1998). *El Nino 1997-98; damaging landslides in the San Francisco Bay area* (USGS Numbered Series No. 089–98; Fact Sheet). U.S. Dept. of the Interior, U.S. Geological Survey, National Landslide Information Center,. http://pubs.er.usgs.gov/publication/fs08998

Hong, H., Xu, C., & Bui, D. T. (2015). Landslide Susceptibility Assessment at the Xiushui Area (China) Using Frequency Ratio Model. *Procedia Earth and Planetary Science*, *15*, 513–517. https://doi.org/10.1016/j.proeps.2015.08.065

Haugerud, Ralph A. (2014), "Preliminary interpretation of pre-2014 landslide deposits in the vicinity of Oso, Washington" (PDF), U.S. Geological Survey, Open-File Report 2014-1065,doi:10.3133/ofr20141065

Iverson, R. M. (2000). Landslide triggering by rain infiltration. *Water Resources Research*, *36*(7), 1897–1910. https://doi.org/10.1029/2000WR900090

Leshchinsky, B. A., Olsen, M. J., & Tanyu, B. F. (2015). Contour Connection Method for automated identification and classification of landslide deposits. *Computers & Geosciences*, *74*, 27–38. https://doi.org/10.1016/j.cageo.2014.10.007

Loehr, J. E., Lutenegger, A., Rosenblad, B., Boeckmann, A., 2017. GEOTECHNICAL SITE CHARACTERIZATION - GEOTECHNICAL ENGINEERING CIRCULAR NO.5. FHWA NHI-16-072

McKean, J., & Roering, J. (2004). Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry. *Geomorphology*, *57*(3–4), 331–351. https://doi.org/10.1016/S0169-555X(03)00164-8

Mwaniki, M. W., Kuria, D. N., Boitt, M. K., & Ngigi, T. G. (2017). Image enhancements of Landsat 8 (OLI) and SAR data for preliminary landslide identification and mapping applied to the central region of Kenya. *Geomorphology*, *282*, 162–175. https://doi.org/10.1016/j.geomorph.2017.01.015

Mathworks, 2019. https://www.mathworks.com/help/images/ref

Ozdemir, A., & Altural, T. (2013). A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. *Journal of Asian Earth Sciences*, *64*, 180–197. https://doi.org/10.1016/j.jseaes.2012.12.014

Pardeshi, S. D., Autade, S. E., & Pardeshi, S. S. (2013). Landslide hazard assessment: Recent trends and techniques. *SpringerPlus*, *2*(1), 523. https://doi.org/10.1186/2193-1801-2-523

Pham, B. T., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA*, *149, Part 1*, 52–63. https://doi.org/10.1016/j.catena.2016.09.007

Pennsylvania Department of Conservation and Natural Resources, 2019. https://www.dcnr.pa.gov/Pages/default.aspx

Pennsylvania Spatial Data Access, 2020. The Pennsylvania Geospatial Data Clearinghouse http://www.pasda.psu.edu/

Rossi, G., Catani, F., Leoni, L., Segoni, S., & Tofani, V. (2013). HIRESSS: A physically based slope stability simulator for HPC applications. *Nat. Hazards Earth Syst. Sci.*, *13*(1), 151–166. https://doi.org/10.5194/nhess-13-151-2013

Saito, H., Nakayama, D., & Matsuyama, H. (2009). Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. *Geomorphology*, *109*(3), 108–121. https://doi.org/10.1016/j.geomorph.2009.02.026

Sato, H. P., Yagi, H., Koarai, M., Iwahashi, J., & Sekiguchi, T. (2007). Airborne LIDAR Data Measurement and Landform Classification Mapping in Tomari-no-tai Landslide Area, Shirakami Mountains, Japan. In K. Sassa, H. Fukuoka, F. Wang, & G. Wang (Eds.), *Progress in Landslide Science* (pp. 237–249). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70965-7_17

Tarboton, D. G., 2003. Rainfall Runoff Processes, Civil and Environmental Engineering Faculty Publications. Paper 2570.https://digitalcommons.usu.edu/cee_facpub/2570

Trigila, A., Iadanza, C., Esposito, C., & Scarascia-Mugnozza, G. (2015). Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology*, *249*(Supplement C), 119–136. https://doi.org/10.1016/j.geomorph.2015.06.001

United State Department of Agriculture, 2019. https://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm

Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., & Hong, H. (2020). Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *CATENA*, *187*, 104396. https://doi.org/10.1016/j.catena.2019.104396

Yeon, Y.-K., Han, J.-G., & Ryu, K. H. (2010). Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Engineering Geology*, *116*(3), 274–283. https://doi.org/10.1016/j.enggeo.2010.09.009

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, *13*(5), 839–856. https://doi.org/10.1007/s10346-015-0614-1

Yatheendradas, S., Kirschbaum, D., Nearing, G. et al. Comput Geosci (2019) 23: 495. https://doi.org/10.1007/s10596-018-9804-y

BIOGRAPHY

Yashar Alimohammadlou received his Bachelor of Science in Civil Engineering from the University of Tabriz, Iran, in 2009 and his Master of Science in Civil engineering with a concentration in Geotechnical engineering from University of Zanjan, Iran in 2012. He joined George Mason University in 2016 for PhD degree in Geotechnical Engineering. He was Graduate Teaching Assistant in the Sid and Reva Dewberry Department of Civil, Environmental and Infrastructure Engineering for three years. He is currently working at DC Department of Transportation.