Data Mining For Knowledge-Based Approach for Landslide Susceptibility Mapping

By

Aaron D. Schuck

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

(Cartography and GIS)

at the

UNIVERSITY OF WISCONSIN – MADISON

2018

Table of Contents

Acknowledgements:

I would like to thank first and foremost my academic advisor, A-Xing Zhu. He helped me bring focus and clarity to not only my work, but my way of thinking. This work could not have been completed without his seemingly limitless patience.

A significant amount of help came from Sharon Khan, our former graduate program director. It was very difficult pursuing a degree while also working full-time, and it would have been much more difficult without her guidance and understanding.

I would also like to thank my committee members, Qunying Huang, and Song Gao. Their feedback helped make this a better product.

My employer, TDS Telecom, provided funding for this endeavor, and was especially patient while I continued to work for them and for the completion of this project.

Lastly, I would like to thank my family. My father was a good example, who also received a master's degree while working full-time. My mother provided ample support and understanding during those times when things became exceedingly difficult. My children, Aleksei and Sophia, without whom this project would be pointless, you both are the love of my life.

List of Figures

List of Tables

Abstract

Statistical approaches to landslide susceptibility mapping can be an effective tool for determining areas of high risk, but have also had problems with portability, reliability, stability, and impracticality as they often require large amounts of data in order to be effective. The knowledge based approach has been proven to address these issues, but the necessary preparation required for obtaining domain knowledge on landslide susceptibility and predisposing factors demands time and effort from the analyst. This paper proposes a knowledge based data mining approach: data mining techniques for defining the knowledge on the relationship between landslide susceptibility and predisposing factors and the knowledge based approach for mapping landslide susceptibility. A set of environmental clusters were generated through fuzzy classification of key environmental layers describing spatial variation of predisposing factors. These clusters were then hardened and later ranked according to density of landslides. The cluster centroid values are ordered in the order of landslide density for hardened classes and used as control points for the construction of fuzzy membership functions on knowledge of relationships between landslide susceptibility and predisposing factors. The fuzzy membership functions are then used through a knowledge-based approach to map landslide susceptibility. The accuracy of the so generated map in the Kaixian area is comparable with that from a logistic regression model. However, the generated fuzzy membership functions are more portable than those in the logistic model. It was also found that the combination layer was correlated strongly with landslide instances, and could serve as a useful input for future studies. The results of this study provide a useful illustration for potential in the construction of fuzzy membership maps from field observation data.

Chapter 1: Introduction

The USGS has conservatively estimated that landslides annually cause between 1-2 billion dollars in property damage within the United States alone. Landslides occur in all 50 states, causing between 25-50 fatalities a year (USGS, 2010). The United Nations has stated "By 2030, urban areas are projected to house 60 per cent of people globally and one in every three people will live in cities with at least half a million inhabitants…Of the 1,692 cities with at least 300,000 inhabitants in 2014, 944 (56 per cent) were at high risk of exposure to at least one of six types of natural disaster (cyclones, floods, droughts, earthquakes, landslides and volcano eruptions), based on evidence on the occurrence of natural disasters over the late twentieth century. Taken together, cities facing high risk of exposure to a natural disaster were home to 1.4 billion people in 2014" (Gu et al., 2015; United Nations, 2016). Numerous studies have recognized human made interactions, including the presence of urban development (roads, buildings, etc.) as one of the leading predisposing factors towards a landslide event (Devkota et al., 2013; Kamp et al., 2008; Yalcin et al., 2011; Youssef et al., 2016).

1.1 Significance

As urban development continues into areas prone to landslides, the need for accurate assessment of landslide susceptibility becomes more and more urgent. GIS is in a unique position to assist in providing landslide prediction and susceptibility mapping. Aside from the actual triggering mechanism for an individual landslide event, all of the underlying predisposing factors that cause landslides can be represented within GIS.

In fact, many studies have already been designed around landslide prediction (Alexander, 2008; Brenning, 2005; Yalcin et al., 2011). Statistical methods are the most widely used for

mapping landslide susceptibility. Logistic regression is the representative of statistical methods. However, logistic regression models have been shown to lack stability, reliability, portability, and scalability (Ercanoglu and Gokceoglu, 2004; Guzzetti et al., 1999; Wang, 2008). Zhu et al. (2014) highlighted the unreliability of the statistical approach:

> "...statistical approaches are often based on linear or generalized linear models, which can only represent the relationships in a monotonic way (inherent generalization; Van Westen et al., 2003). However, the actual relationships are complex and inherently nonlinear. For example, the relationship between strata (strike and dip) and landslide susceptibility is highly nonlinear because this relationship is also related to the slope information including gradient and aspect (Atkinson and Massari, 1998; Donati and Turrini, 2002; Lee et al., 2002; Liu et al., 2004; Zhu et al, 2004; Ayalew and Yamagishi, 2005). These linear or generalized linear models are thus insufficient to represent complicated nonlinear relationships."

The knowledge-based approach was designed to address these deficiencies and has demonstrated its ability to function well in those areas where traditional statistical approaches have failed. Zhu et al. (2014) provide the following characterization for the knowledge-based approach:

> "The results of our case studies suggest that this knowledge-based approach holds up well when it is transferred without changes to an area that is about 19 times larger and much more complicated than the area in which the knowledge base was developed... The expert knowledge approach in this study does not use past landslides to develop the knowledge base. It is essentially different from the statistical methods in that the expert knowledge approach does not use data of landslide occurrence and absence to extract the relationships

between landslide susceptibility and predisposing factors. It does not have the false negatives during its model development as the statistical methods and some of the data mining methods do."

Though the knowledge-based approach has been proven to be a more accurate method for mapping landslide susceptibility (Wang, 2008; Zhu et al., 2014), it does come with its own disadvantages. The knowledge based-approach is reliant on extracting domain knowledge through extensive and detailed interviews with landslide experts. Experts provide a detailed description for each pre-disposing factor within the landscape and how that pre-disposing factor will contribute to landslides. This knowledge is then assimilated into the computer through the form of a fuzzy membership function for each pre-disposing factor. Each pixel will then have the pre-disposing factors aggregated, and a landslide susceptibility score is generated for every pixel over the landscape by integrating the values of predisposing factors and the fuzzy membership functions.

The fuzzy membership functions are the key to the knowledge based approach (Yang et al., 2011; Zhu et al., 2010). While the landslide susceptibility map is the desired result, those who wish to process and analyze landslide susceptibility for a given area may not have the time, the resources, or the experts to conduct these interviews and construct the necessary fuzzy membership functions needed for the knowledge based approach to work. Very often all an analyst has to work with is existing observations of landslide occurrences.

1.2 Research Question

In order to address the challenges in defining membership functions for knowledge-based approach, this thesis explores the following scientific question: Is it possible to construct fuzzy membership functions from a sample data-set only, bypassing the need for extensive knowledge-based interviews?

1.3 Goal of Thesis

The ultimate goal of this thesis is to provide an alternative method for obtaining the knowledge for the knowledge based approach so that practitioners can use the benefits of the knowledge based approach, whilst simultaneously avoiding the drawbacks of the statistical approaches. The approach described within this work aims to be less costly in terms of the necessary time and effort that must be expended when applying the knowledge based approach to a given area. Given a dataset, this methodology will provide a more reliable result than seen with standard statistical approaches while also eliminating the prerequisites for the knowledge based approach.

1.4 Thesis Structure

The remaining portion of the thesis will expand on the reasoning behind why building this new approach is necessary, and then continue to describe the construction of the methodology used to attain the desired results. Chapter 2 will cover the challenges of implementing the statistical methods for landslide susceptibility mapping effectively, and provide a background on the knowledge based approach. Chapter 3 will present an overview of the Kaixian and Three Gorges study areas within the first subsection. The second subsection will

describe the methodology of this approach. The fourth chapter will provide an overview of the results, and the fifth chapter will conclude this thesis and present avenues for future research direction.

Chapter 2: Background on Knowledge Based Approach

2.1 Challenges of the statistical methods for landslide mapping

Multi-variate statistical approaches have been the predominant method used to map landslide susceptibility, and there are several variations to the approach (Bai et al., 2010; Lee et al., 2005, 2006; Nefeslioglu et al., 2008; Yilmaz 2009, 2010). Albeit their wide application, the statistical approaches suffer from the following drawbacks (Zhu et al., 2014):

1) Statistical methods use the landscape characteristics of the sample sites (landslide occurrence sites as positive evidence and non-landslide sites as negative evidence) to define the relationships between landscape and predisposing factors (Zhu et al., 2014). The quality of landslide data has uncertainty built into it because of the very nature of the phenomenon. Even experts may disagree as to the exact shape and area of a landslide occurrence (Atkinson and Massari, 1998). The predisposing factors for a given landslide will change after a landslide event, often to such a degree that completely new values are present at a landslide site after each event, adding to the difficulty of obtaining accurate data (Guzzetti et al., 1999). Such ambiguity manifests as uncertainty within our data set, and eventually in the algorithm used to build the logistic regression model.

2) Statistical methods have historically shown low stability. Data driven models are very sensitive to their training datasets. Unique coefficients are calculated based on the data for a given sample set, and no two sample sets will have the same values. As such, unique coefficients lead to unique results, and different equations. This situation will often lead to results that conflict with reality, or one another. When conducting a purely data driven approach,

ignoring the physical processes and causes of a landslide event are critical omissions which can lead to misleading results (Dai et al., 2002).

For example, Wang (2008) conducted experiments using the logistic regression model using landslide data from China. Two experiments were conducted with the same data. Experiment A trained the logistic regression model using 21 pixels where landslides had occurred, and 21 pixels where landslides had not occurred. Experiment B did the same, except the 21 negative pixels were instead determined at random. The results from Experiment A suggested that 4 predisposing factors were the most important, while the results of experiment B suggested 3. In addition, factors which have been determined as very important indicators of landslide susceptibility by experts were instead negatively weighted for the model (Wang, 2008). The lack of consistency amongst the tests demonstrates the low stability inherent within this approach.

3) Statistical methods are not portable. It has been discussed that the training data shapes the algorithm for data driven models. Every area has a different algorithm to go with its own unique data, and a statistical model is built tailored to the region. Because of this, statistical models used for one area can't be extrapolated to surrounding areas (Guzzetti et al., 1999; Carrara et al., 1991), resulting in a lack of portability.

4) It is impractical to use statistical methods for large areas. It is well known that data-driven models and statistical techniques require large amounts of data to produce reliable results (Ercanglou, 2004). The larger the study area, the more samples are needed, and with budget and time constraints, this may not always be feasible.

2.2 Knowledge-based approach on landslide susceptibility mapping

The knowledge based approach is designed in such a way that it can effectively address most of the problems we see within the purely statistical methods (Zhu et al., 2014). Figure 2.1 outlines the basic idea of the knowledge based approach. Interviews are conducted with local landslide experts to determine which predisposing factors will be included in the calculation of landslide susceptibility and how these predisposing factors affect landslide susceptibility. This knowledge is captured and represented as fuzzy membership functions for each predisposing factor. Based on these fuzzy membership functions together with the values of the predisposing factors, every individual pixel within the study area is then evaluated to derive the landslide susceptibility; 0 being not susceptible at all to landslides based on this predisposing factor, 1 being as susceptible to landslides as possible based on this predisposing factor.

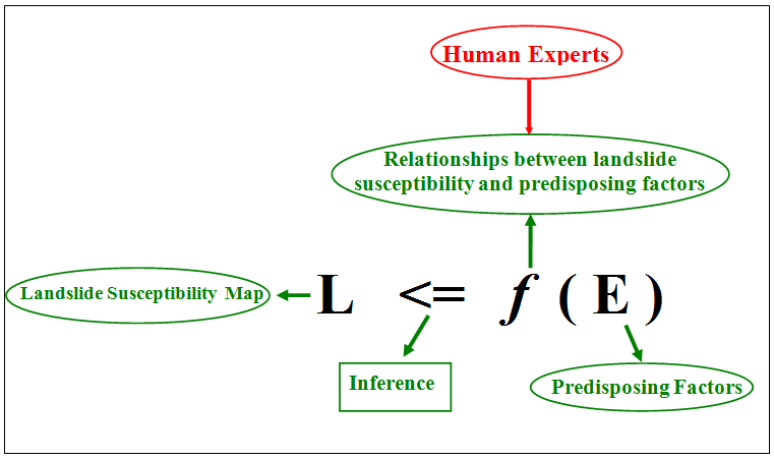

Figure 2.1: Basic diagram of knowledge based landslide susceptibility mapping approach. (Taken from Wang, 2008)

This approach has the following advantages over purely statistical methods:

1) Expert knowledge is built directly into the model. Instead of relying purely on the sample data to shape the statistical relationships as it is done in the commonly used logistic

regression, expert knowledge on physical reasons for landslide susceptibility is incorporated into the equation, resulting in greater reliability and consistency.

2) The knowledge based approach can be used in different areas. The relationships between landslides and their predisposing factors will be extracted directly from the experts. Expert knowledge is developed over years of observing landslide phenomenon and working in the field to understand the underlying factors which cause them. Using this knowledge as a foundation, the knowledge based approach is able to be used as a generic model for a variety of areas across the globe, instead of being generated on the fly using local sample data on a case by case basis as we have seen with existing statistical methods. This allows the same relationships to be used in a variety of areas. This portability also means that the knowledge based approach is less data hungry, allowing larger areas to be mapped.

3) The knowledge based approach is not restricted to linear models. The relationship between a given predisposing factor and the probability of a landslide occurrence is rarely, if ever, linear (Zhu et al., 2014). The knowledge based approach designs the curves for the relationships around each predisposing factor and its landslide probability, thus avoiding any linearity assumptions.

While the knowledge based approach addresses most of the deficiencies of the statistical methods, there is one major concern that should be examined. The knowledge based approach assumes that detailed knowledge extraction will take place. Interviews will be conducted, curves will be identified and shaped to each individual predisposing factor based on the expert's knowledge, and the fuzzy membership equations will be constructed based on the information gained from the expert(s).

It will often be the case however, that the analyst will not have the time or the resources to conduct the thorough investigation that is required in order for the knowledge based approach to be successful. All that they will have is a dataset of landslide occurrences to work with. In this type of situation, the most must be made from the dataset alone. The question then is: how to build fuzzy membership functions from these landslide occurrence data.

Chapter 3: Materials and Methods

3.1 Study areas

Two sites were chosen for the study area: Kaixian and the adjacent Three Gorges area (Figures 3.1-3.2).

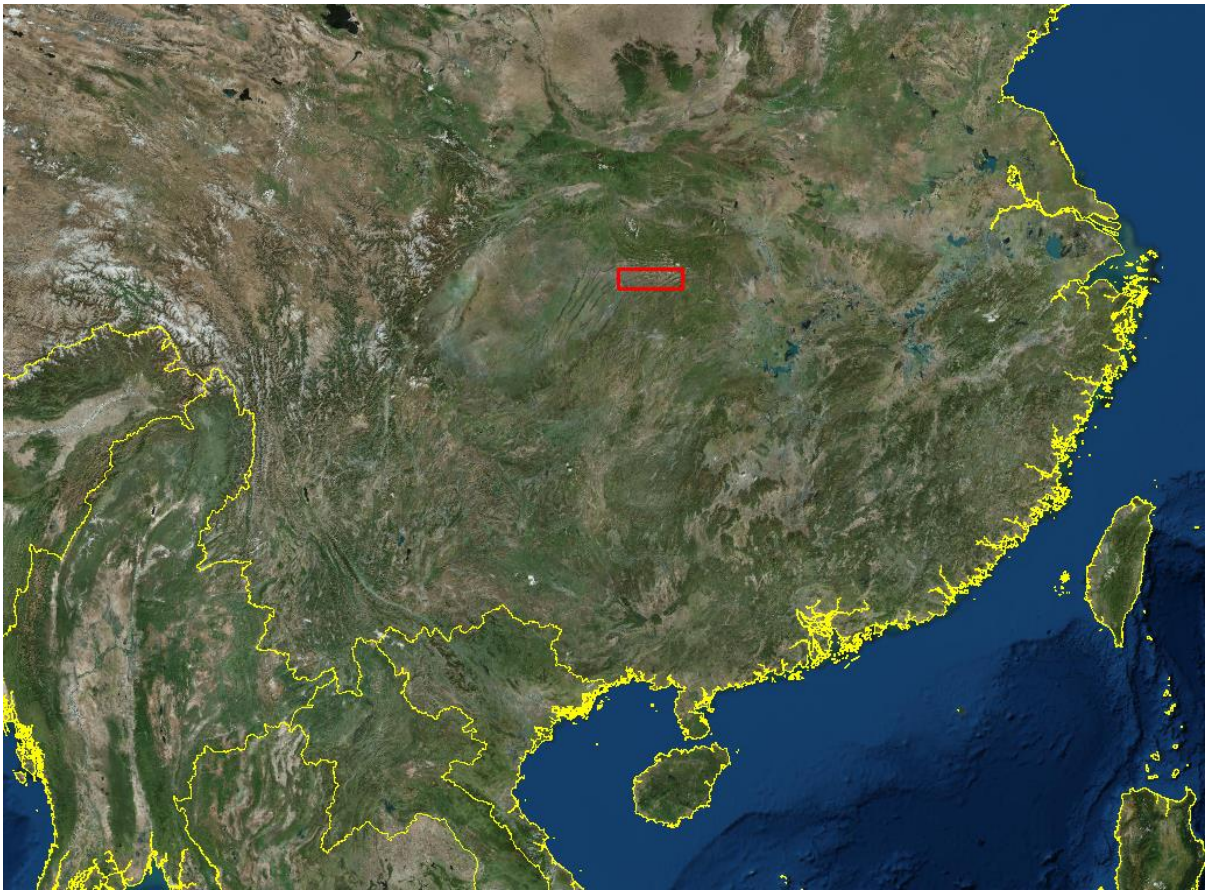

Figure 3.1: The Kaixian and Three Gorges study areas are located within the red bounding box.

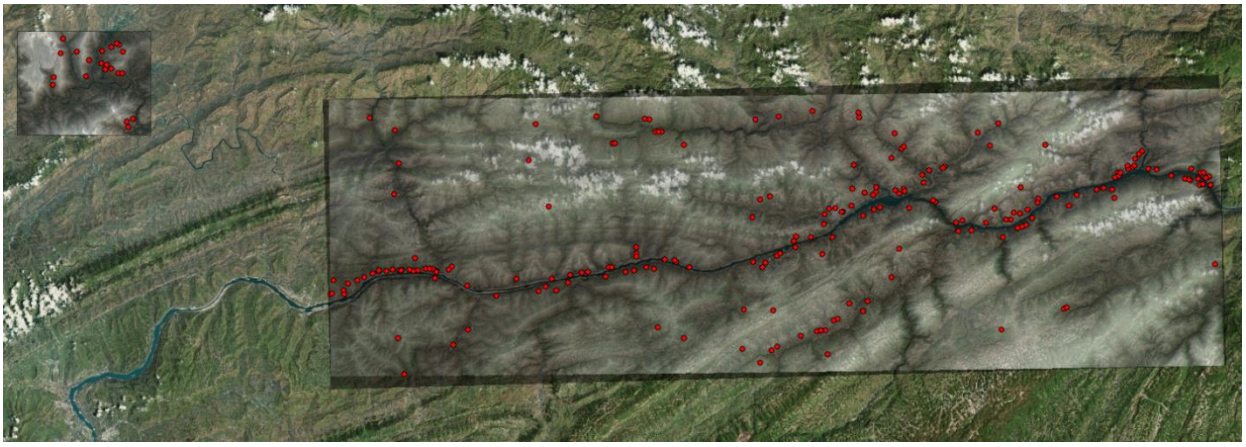

Figure 3.2: The Kaixian and Three Gorges study areas. DEM's are overlaid the study areas, recorded landslides are displayed in red.

There are a total of 226 recorded landslide events, 21 within Kaixian, and the remaining 205 within the Three Gorges area. The Kaixian study area is focused around the town of Kaixian County with an area of 250 square kilometers (18 km * 14 km). The Nanhe river passes through the town, surrounded by steep cliffs. The greatest local relief in this area is approximately 700 meters, with an average slope gradient of roughly 10 degrees. The Three Gorges area is considerably larger with an area of 4,440 square kilometers (37 km * 120 km). The Yangtze River bisects the entire area. The area has an average slope gradient of approximately 32, with the largest local relief being 1,671 meters. The Kaixian area was used to develop fuzzy membership functions which were then applied to the Three Gorges area in order to test the portability of the methodology.

3.2 Methodology

To extract the membership functions relating landslide susceptibility to environmental variables (predisposing factors) from landslide observation datasets, a set of environmental clusters were created using the FCM method on the given environmental variables for the study

area. The environmental clusters were then hardened to create a map showing the core areas of these clusters. The landslide instances were overlaid onto the hardened map. The clusters were then ranked according to their landslide density. Fuzzy membership functions were then constructed using the centroids of the ranked clusters and used together with the environmental variables to predict landslide susceptibility.

For a given variable, the landslide susceptibility is at its highest for that variable when its value is equal to the cluster centroid of the cluster with the highest landslide density. The susceptibility at this environmental value will then be set to 1 for the fuzzy membership function curve (i.e. the most susceptible to landslides for the given value for this variable). The susceptibility is 0 for the environmental value of this variable when the environmental value is equal to the centroid of the cluster with the lowest landslide susceptibility. The remaining cluster values filled in the curve depending on where they ranked in terms of landslide density. This process was repeated for each variable, resulting in a set of fuzzy membership functions. The final step used SoLIM Solutions software to create a landslide susceptibility map using the constructed membership functions and the spatial variation of the predisposing factors.

3.2.1 Selection of Pre-Disposing Factors

The details of constructing the knowledge based approach have been well documented (Wang, 2008; Zhu et al., 1994, 2001). Wang (2008) worked with a local landslide expert who identified seven essential predisposing factors which play a crucial role in determining landslide susceptibility in the Kaixian study area: lithology (rock type), strata dip, strata strike, slope gradient, slope aspect, slope relative relief, and slope shape.

Initial attempts at clustering with the seven layers proved problematic. For one, categorical values such as lithology and slope shape cannot be meaningfully clustered with data on the ratio scale without separating the dataset into several different categories, and then further clustering within these categories.

Zhu et al. (2014) successfully integrated several key environmental layers through the following formula:

$$f_{Strata\ Slope\ (d_{ij},\, s_{ij}, g_{ij}, a_{ij})} = \begin{cases} 0.0 & if\ |s_{ij} - a_{ij}| > 90 \\ 0.0 & if\ d_{ij} > g_{ij} \\ \exp\left(-\left(\frac{|d_{ij} - g_{ij}| \text{ x } 0.8326}{45}\right)^2\right) \text{ x } \cos(s_{ij} - a_{ij}) & \text{otherwise} \end{cases}$$

(3.1)

where d_{ij} is the strata dip at cell (i,j); s_{ij} is the strata strike at cell (i,j); $g_{ij}$is the slope gradient at cell (i,j); and a_{ij} is the slope aspect at cell (i,j).

An issue arises when calculating the cosine portion for this equation, namely there are instances where the slope aspect for 2 given pixels will be quite similar but will yield numerically dissimilar results. For example, if we take two pixels, one with a slope aspect of 359° and a strata strike of 300° and its neighbor with a slope aspect of 10° with the same strata strike of 300° and insert them into the cosine portion of formula (1) we will get .515 & .342 respectively (See figure 3.3).

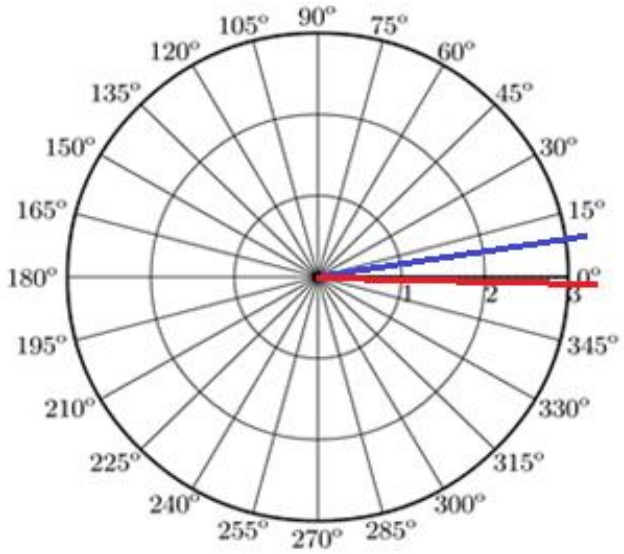

Figure 3.3: The red line (359 ^{o}) and the blue line(10 ^{o}) are both facing in approximately the same direction, but not accounting for the disparity in values can cause errors.

To counteract this, the following was substituted for the $(s_{ij} - a_{ij})$ portion of Equation 3.1:

$$\begin{cases} s_{ij} + (360 - a_{ij}) & if\, s_{ij} \leq 90\ and\ a_{ij} \geq 270 \\ a_{ij} + (360 - s_{ij}) & if\, a_{ij} \leq 90\ and s_{ij} \geq 270 \\ |s_{ij} - a_{ij}| & \text{otherwise} \end{cases}$$

(3.2)

Equations 3.1 & 3.2 were used to generate an environmental layer, hereafter referred to as the combination layer. This combination layer was then joined together with the slope gradient and slope height layers as the inputs for the FCM clustering process.

3.2.2 Generation of Fuzzy Clusters

Several studies have successfully predicted new terrain characteristics through generating clusters of known environmental attributes (De Bruin and Stein, 1998; Yang et al., 2011; Zhu et al., 2010). The FCM utility within the SoLIM program was used to generate cluster maps, along with cluster centroid values for each of the contributing data inputs.

Fuzzy c-means clustering is an unsupervised classification technique that detects clusters of the provided variables through iteratively identifying cluster centroids (Bezdek et al., 1984). Individual pixels are assigned a degree of membership to every cluster based on their distance from the cluster centroids and their degree of belonging to that given cluster. The weighting exponent (*m*) and the number of clusters (*c*) are the two important parameters for FCM clustering.

The *m* value controls the weights assigned to the distance from a given pixel to a centroid. The larger the *m* value, the fuzzier the pixels will be (every pixel belongs to every cluster). If the *m* value is smaller, the clusters will become less fuzzy. This research assigned *m* = 2, as it has been proven to give good clustering results (Pal and Bezdek, 1995).

The *c* value controls the number of clusters to be generated. Determining the optimum number of clusters to generate is a challenging and well known problem (Pal and Bezdek, 1995; Wang and Zhang, 2007; Xu and Wunsch, 2005). Extensive research has been devoted to cluster validity indices, which are designed to determine the optimum number of clusters. The partition coefficient and the entropy values are among the most popular indices used to determine cluster validity (Bezdek et al., 1984). This research used these indices to select the best cluster fit.

3.2.3 Generation of Fuzzy Membership Functions

The generated environmental clusters were then hardened. Any pixel containing a membership ≥ 0.5 in a cluster will be treated as a representative of that cluster (setting its value to 1); the remaining pixels for that cluster were discarded (setting their values to a value of 0). The landslide instances were overlaid onto the hardened cluster maps. Landslide density per square kilometer was calculated for each environmental cluster. The generated clusters were then ranked in terms of landslide density. Clusters with the highest landslide density were categorized as high landslide susceptibility environments, and the reverse was true for areas with low landslide density.

Two types of knowledge are needed in order to define a fuzzy membership function (Zhu, 1999). Type 1 knowledge consists of the typical environmental conditions present for a given landslide susceptibility. The FCM clustering process generated this knowledge using the cluster centroid values for the environmental clusters with highest landslide density. Type 2 knowledge states how the membership changes when the environmental condition deviated from its value for the highest susceptibility. This type of knowledge is determined by the centroids of the clusters with less landslide density. The cluster centroid values were plotted for each variable listed. With sufficient cluster centroids, a curve can be constructed which will form the fuzzy membership function (See Figure 3.4). This process was repeated for each of the three variables identified. The fuzzy membership curves for all of the environmental variables are then created.

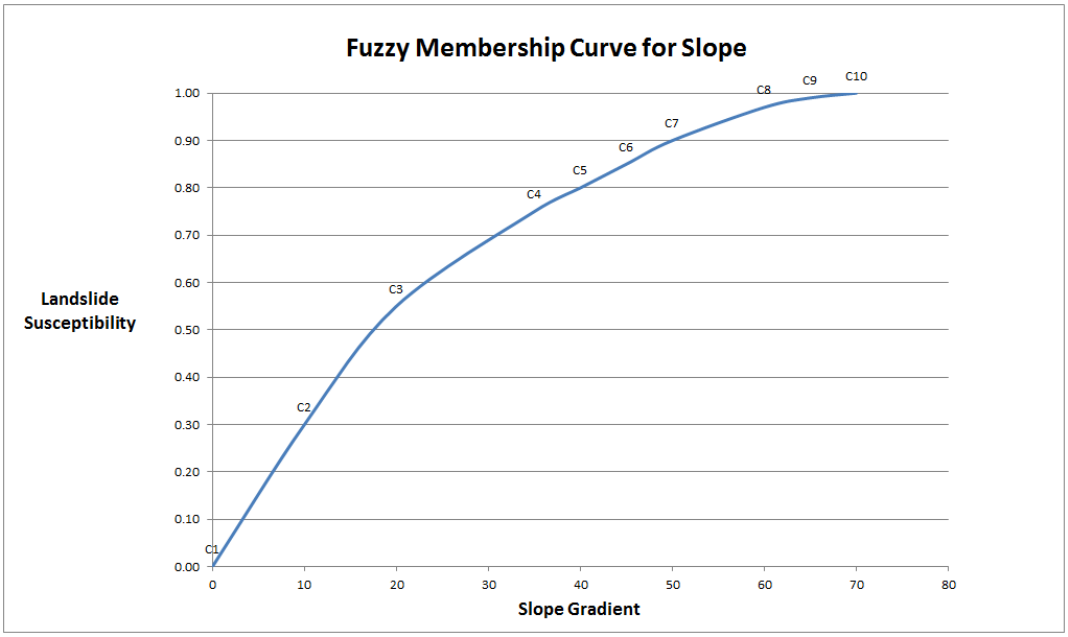

Figure 3.4: An example fuzzy membership curve.
The C values shown are the cluster centroids generated in fuzzy classification.

3.2.4 Computation of Landslide Susceptibility

The inference engine in the SoLIM Solutions software was used to generate the landslide susceptibility maps. A rule-based project was constructed, and the 3 data layers served as input into the GIS database. Landslide susceptibility served as a "soil instance", and each of the 3 layers had s-shaped fuzzy membership curves defined mirroring the graphed data built around the centroid values of each of the clusters. The inference kept the same 5 meter resolution as each of the input layers. Areas containing 'NoData' values were masked out to produce reliable results.

Chapter 4: Results and Discussion

4.1 Environmental Clusters

Each distinct number of cluster datasets produces a partition coefficient and an entropy level. Figure 4.1 highlights the results from the clustering process. A low entropy value implies a more "crisp" partition between clusters, while a high partition coefficient value implies a least fuzzy clustering amongst the different cluster sets (Bezdek, 1974). Cluster sets 2 through 6 have the lowest entropy and the highest partition coefficient. The 3 cluster dataset was chosen because it correlated well with the desired scaling for a landslide susceptibility map (A cluster for areas with low, medium, and high landslide susceptibility).

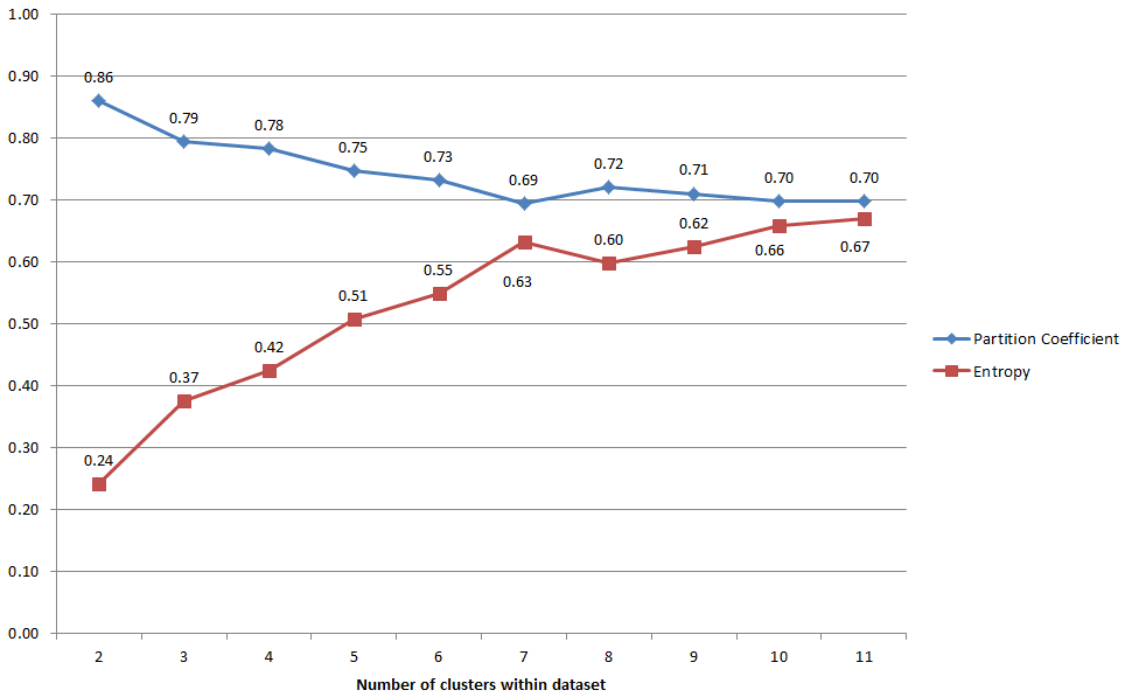

Figure 4.1: Initial data point output of the FCM process.

After the clusters were generated, the clusters were hardened by assigning a category to every pixel based on its score (Figure 4.2).

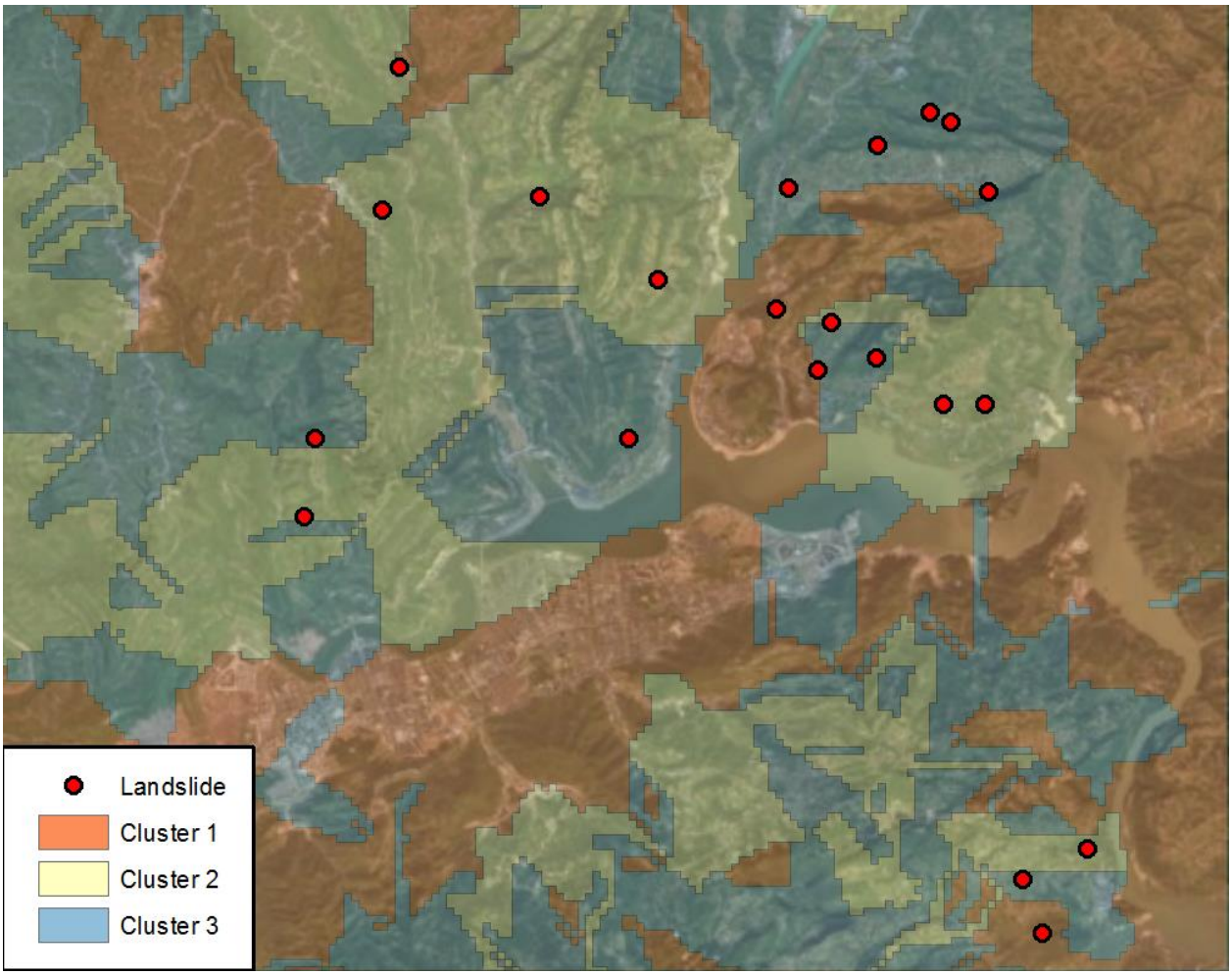

Figure 4.2: 3 cluster dataset overlaid on the city of Kaixian with landslide data-points.Cluster 1 is red, cluster 2 is yellow, and cluster 3 is blue.

4.2 Fuzzy Membership Functions

The variables were then graphed to facilitate construction of fuzzy membership functions. The centroids of the ranked clusters are shown in Table 4.1. All of the variables produced a graph shape roughly equivalent to an S-shaped curve. Clusters 2 & 3 were similar in terms of

their landslide density. In order to generate a continuous fuzzy membership function, the centroid values of Cluster 3 were assumed to be the most optimal value at which a landslide would occur, while the centroids of Cluster 1 were used as the cross value, or mid-optimal value.

Table 4.1: Resulting clusters arranged by landslide density.

Cluster Number	Combination Layer	Slope Gradient	Slope Height	KM²	Landslides	Density
1	0.0619	5.56	148.6	87.08	3	0.03445
2	0.1422	12.78	490.05	61.27	8	0.13057
3	0.1097	10.11	331.64	72.93	10	0.13712

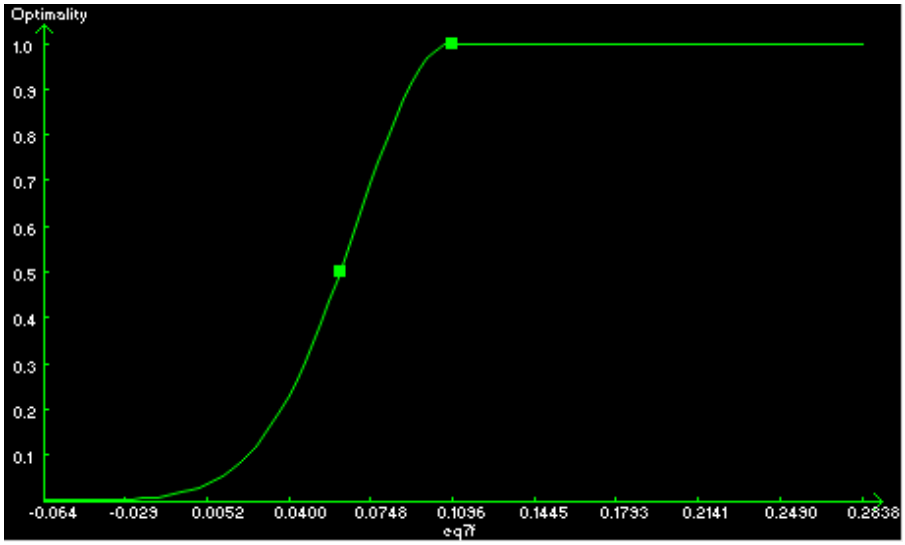

Figure 4.3: The variables correlating to higher landslide density clusters are translated into a SoLIM fuzzy membership function.

4.3 Inferred Landslide Susceptibility Map for Kaixian

SoLIM takes the constructed fuzzy membership curves as input to infer a landslide susceptibility map. A binary map is generated from the inference with a cutoff of 50% membership, any pixel with a susceptibility value greater than or equal to 50% is considered a landslide susceptible area. The resulting landslide susceptibility map for the Kaixian area (Figure 4.4) has a success rate of 66.67%, capturing 14/21 of the landslide sample points. The generated combination layer had the greatest influence on the susceptibility map, and was a strong predictor of the presence or absence of a landslide.

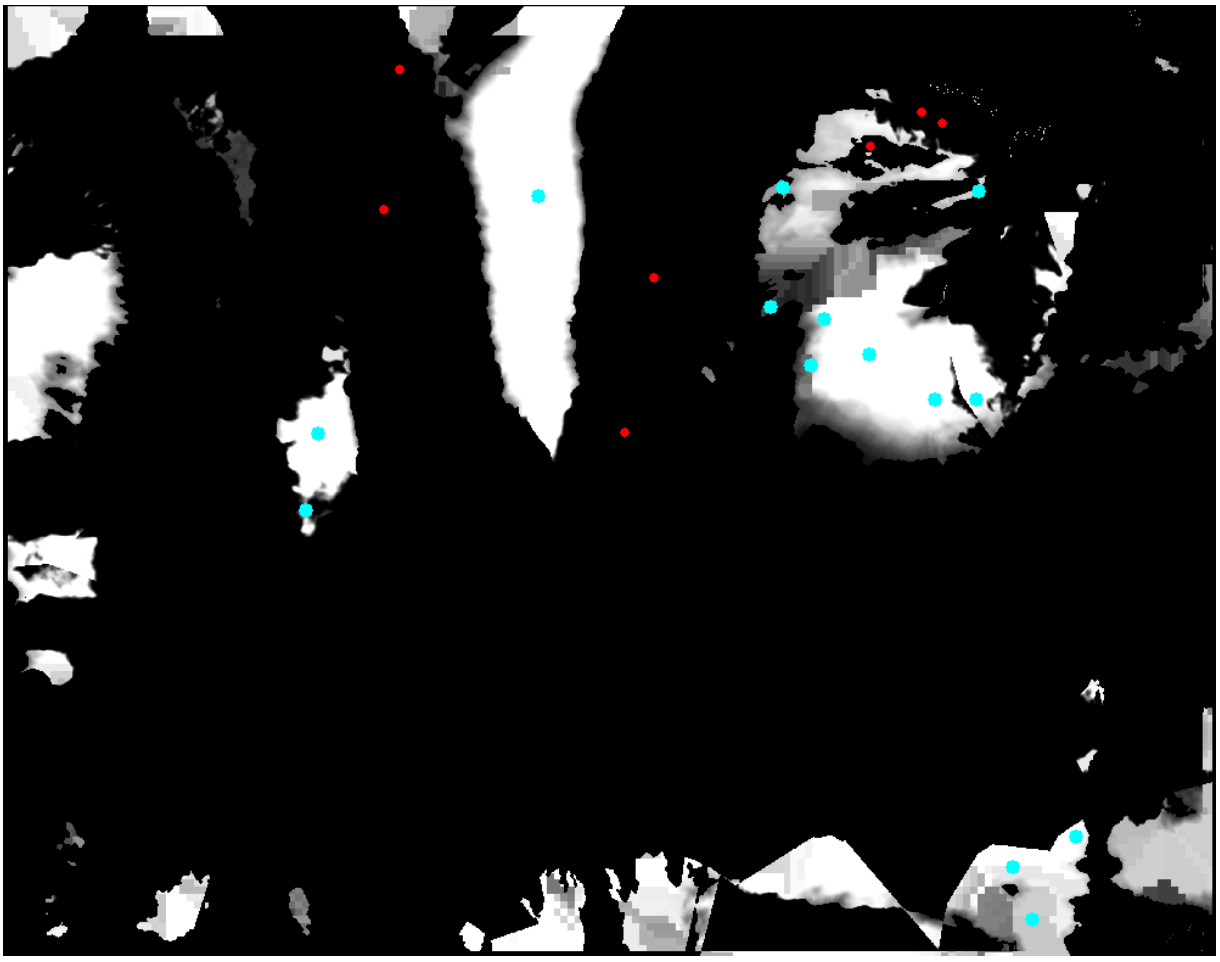

Figure 4.4: The inferred landslide susceptibility map for Kaixian. The white areas are ≥ 50% susceptible to a landslide event. Red dots are outside of landslide susceptible areas, blue dots are within landslide susceptible areas.

4.4 Inferred Landslide Susceptibility Map for Three Gorges

To test the robustness of this methodology, the fuzzy membership curves derived from the Kaixian area were applied to the larger Three Gorges area (Figure 4.5). Selecting those landslides within ≥ 50% susceptible areas captured 74/205 landslides, a success rate of roughly 36%. The lower success rate indicates that the method is not as easily transferable as the knowledge based approach, but we do see again that the generated combination layer is a much more prominent influence than either the included slope gradient or slope height layers.

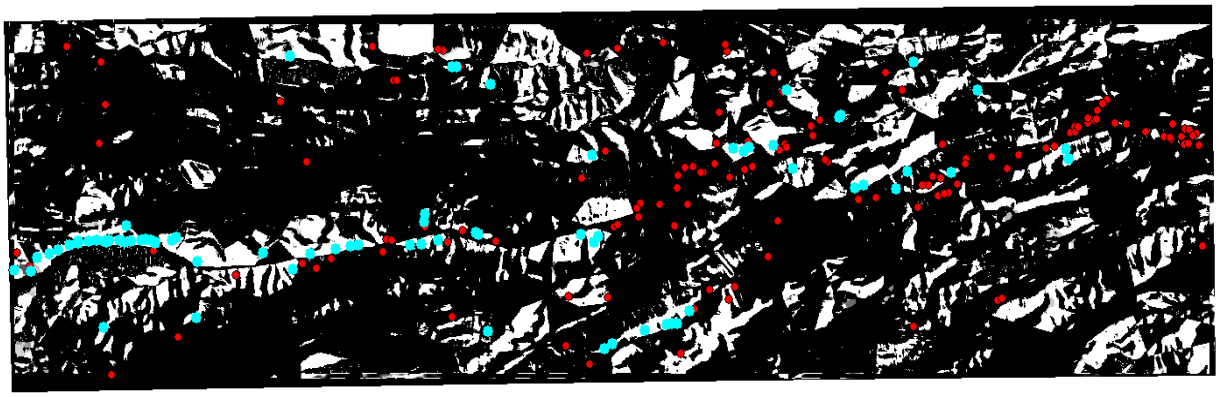

Figure 4.5: The inferred landslide susceptibility map for Three Gorges using fuzzy membership functions derived from Kaixian.

4.5 Logistic Regression Comparison

Despite the problems previously listed concerning the established statistical methods, it is useful in this case to compare a logistic regression output with the proposed methodology. This is mainly because the statistical methods are well understood, and can provide a benchmark to compare how successful the proposed methodology is in achieving its goals.

Unique random samples of 21 non landslide pixels were generated for each trial to train the logistic regression algorithm. Buffers were created around existing landslides and each randomly generated sample point to ensure no sample point could be generated within half a

kilometer from each other, or an existing landslide point. The samples were created using ArcGIS's CreateRandomPoints_management function, and the logistic regression itself was performed using Python, Pandas, and the Statsmodel APIs.

Five trials of logistic regression were run on the Kaixian data (Table 4.2). In general, pseudo R^2 values greater than .2 indicate a good fit (Clark and Hosking, 1986). Trial 3 had the lowest standard error for the combination layer and a pseudo r-squared value of .2193, so it was chosen to create a susceptibility map (Figure 4.6). The resulting map was then classified into areas with low, medium, and high chance of landslide susceptibility using Jenks natural breaks classification method, each category capturing 0, 7, and 14 landslide points respectively.

Table 4.2: The output data for the Kaixian logistic regression trials.

		Coefficient	Std. Error	Z	P>|z|	.025 CI	.975 CI
Trial 1 *Pseudo R-squared: .5315*	Slope Height	0.008	0.005	1.678	0.093	-0.001	0.017
	Slope Gradient	0.055	0.101	0.545	0.586	-0.143	0.253
	Combo Layer	12.4479	9.486	1.312	0.189	-6.144	31.04
	Constant	-4.733	1.916	-2.47	0.014	-8.489	-0.977
Trial 2 *Pseudo R-squared: .4708*	Slope Height	0.0093	0.005	1.962	0.05	1.17E-05	0.018
	Slope Gradient	0.0232	0.071	0.328	0.743	-0.115	0.162
	Combo Layer	7.9798	2.961	2.695	0.007	2.176	13.784
	Constant	-4.9121	2.075	-2.367	0.018	-8.979	-0.845
Trial 3 *Pseudo R-squared: .2193*	Slope Height	0.0041	0.003	1.399	0.162	-0.002	0.01
	Slope Gradient	0.0695	0.079	0.883	0.377	-0.085	0.224
	Combo Layer	2.874	1.069	2.689	0.007	0.779	4.968
	Constant	-3.0043	1.21	-2.482	0.013	-5.376	-0.632
Trial 4 *Pseudo R-squared: .1745*	Slope Height	0.0023	0.003	0.775	0.438	-0.004	0.008
	Slope Gradient	-0.0071	0.052	-0.135	0.892	-0.11	0.095
	Combo Layer	3.0678	1.116	2.749	0.006	0.88	5.255
	Constant	-1.5545	1.117	-1.391	0.164	-3.744	0.635
Trial 5 *Pseudo R-squared: .3139*	Slope Height	0.0033	0.004	0.853	0.393	-0.004	0.011
	Slope Gradient	0.0545	0.087	0.627	0.531	-0.116	0.225
	Combo Layer	5.1258	1.681	3.05	0.002	1.832	8.42
	Constant	-2.9393	1.595	-1.843	0.065	-6.065	0.186

When the map is restricted to a binary output, counting only areas of high landslide susceptibility or no susceptibility, the output is essentially the same as was seen with the proposed methodology (14/21 landslide captured). This suggests that the proposed methodology is at least on par with logistic regression for determining landslide susceptibility.

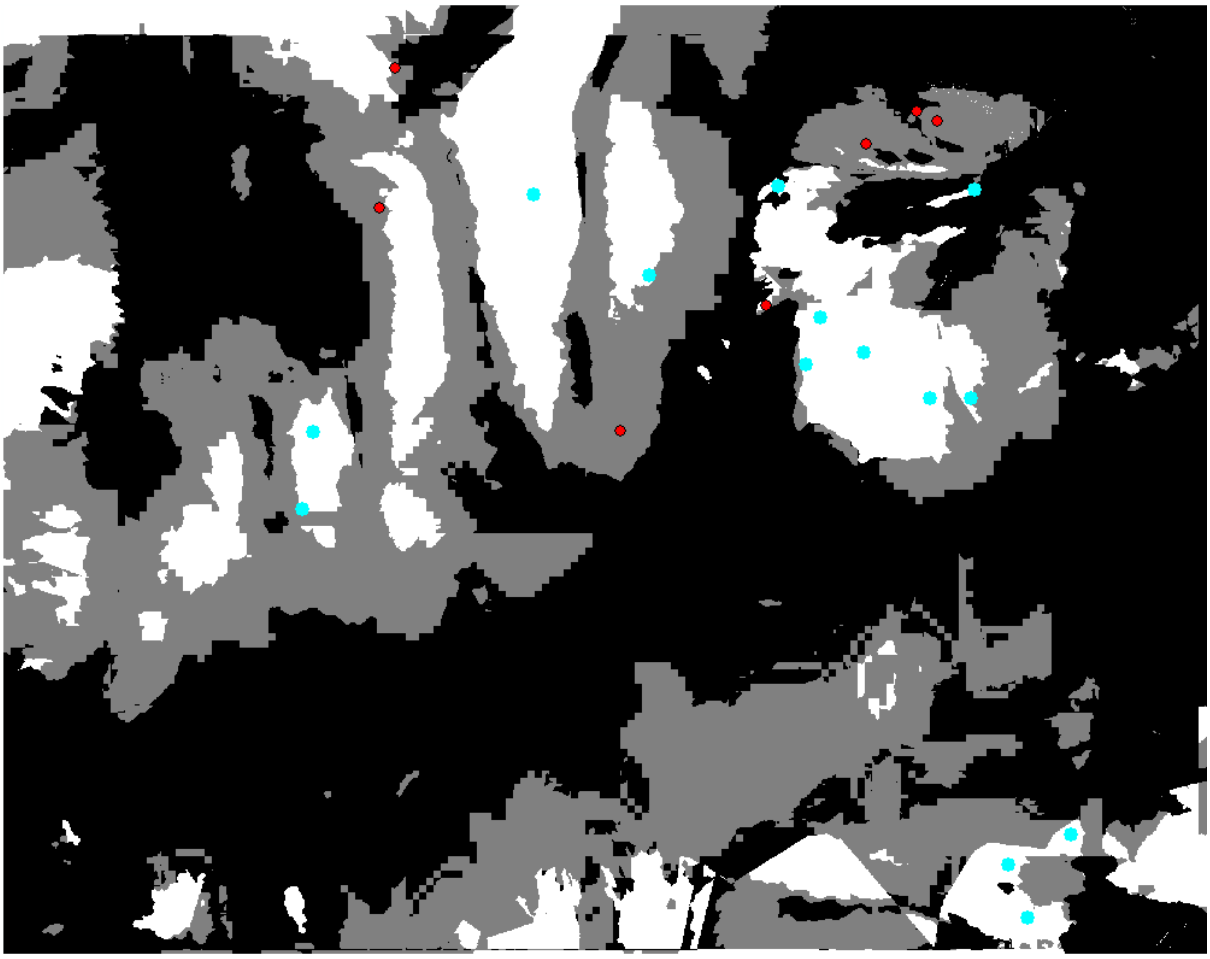

Figure 4.6: The landslide susceptibility map created using logistic regression for Trial 3. The blue dots are those landslides within highly susceptible areas; the red dots are within areas of medium susceptibility.

As was seen with the previous maps, it's interesting to note the generated combination layer was consistently the most heavily weighted out of the three variables, and there was a very

high confidence in the link between its presence and the presence of a landslide. Using this layer alone could yield decent susceptibility maps. Considering the ease with which this layer can be generated, this may be a way to create quick landslide susceptibility maps by mining existing data layers, though further refinement of the technique is required.

The logistic regression coefficients generated in trial three from the Kaixian area were then applied to the Three Gorges area to create a landslide susceptibility map (Figure 4.7). The resulting map captured 69/205 landslides, for an accuracy of 33.65%. This is slightly less accurate than the knowledge based data mining method (See tables 4.3-4.4).

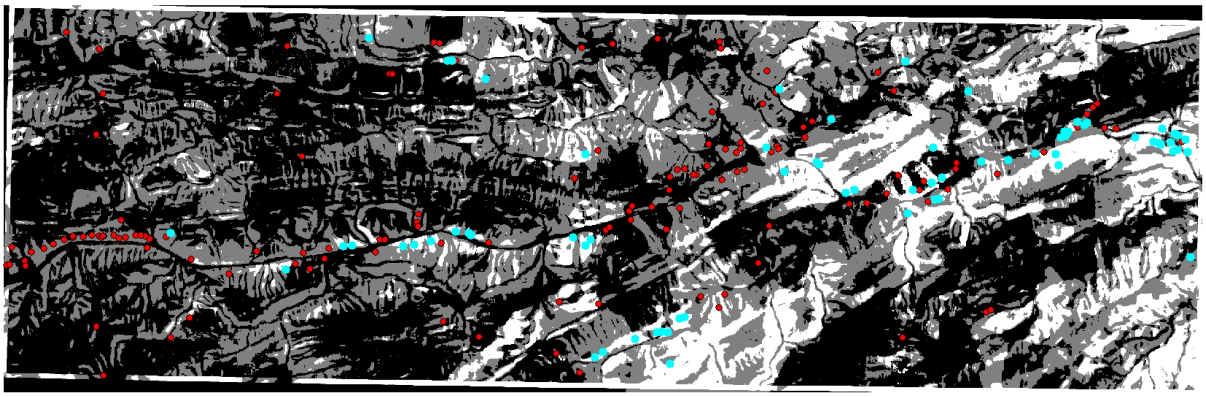

Figure 4.7: Landslide susceptibility map for the Three Gorges area using the logistic regression coefficients from Kaixian. The blue dots are those landslides within highly susceptible areas

Table 4.3: Accuracy comparison for the Kaixian area.

Method	Landslides Captured	Accuracy Percentage
Logistic Regression	14/21	66.66%
Knowledge Based Data Mining	14/21	66.66%

Table 4.4: Accuracy comparison for the Three Gorges area

Method	Landslides Captured	Accuracy
Logistic Regression	69/205	33.65%
Knowledge Based Data Mining	74/205	36.09%

Chapter 5: Conclusion and Future Direction

5.1 Conclusion

The main goal of this research was to determine if fuzzy membership functions could be built using landslide observation data in lieu of the knowledge based approach to reliably determine areas which are highly susceptible to landslides. A methodology was developed that distilled the available data layers into a combination layer which incorporated several of the key environmental variables into one formula. This layer combined with slope gradient and slope height served as inputs into the FCM process. The output clusters from the FCM process were then ranked according to landslide density, and their ranked centroid values served as optimality points to build fuzzy membership functions.

The initial results demonstrated that data mining can be used to identify clusters of higher landslides incidents and construct a suitable landslide susceptibility map. The accuracy in the model development area is comparable with that from a logistic regression model. However, the approach achieves a slightly higher accuracy when it is ported to the Three Gorges area in comparison with the logistic regression model. This implies there is validity in extracting knowledge from the landslide occurrence data for knowledge-based approach to landslide susceptibility mapping.

5.2 Future Direction

The research was conducted with limited data (21 landslide occurrences in the Kaixian area). Future work could test this approach for areas with large sample data size and with more diversity of environmental settings.

The use of limited number of environmental predisposing factors may also contribute to the low accuracy of the new approach. Further work should develop other means to encoding predisposing data to add more dimensions in ability to characterize environments which leads to landslide occurrences.

References:

Alexander, David E. "A brief survey of GIS in mass-movement studies, with reflections on theory and methods." Geomorphology 94, no. 3-4 (2008): 261-267.

Atkinson, Peter M., and R. Massari. "Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy." *Computers & Geosciences* 24, no. 4 (1998): 373-385.

Bai, Shi-Biao, Jian Wang, Guo-Nian Lü, Ping-Gen Zhou, Sheng-Shan Hou, and Su-Ning Xu. "GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China." *Geomorphology* 115, no. 1-2 (2010): 23-31.

Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences* 10, no. 2-3 (1984): 191-203.

Bezdek, James C. "Numerical taxonomy with fuzzy sets." *Journal of Mathematical Biology* 1, no. 1 (1974): 57-71.

Brenning, A. "Spatial prediction models for landslide hazards: review, comparison and evaluation." *Natural Hazards and Earth System Science* 5, no. 6 (2005): 853-862.

Carrara, A., M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Reichenbach. "GIS techniques and statistical models in evaluating landslide hazard." *Earth surface processes and landforms* 16, no. 5 (1991): 427-445.

Clark, William AV, and Peter L. Hosking. Statistical methods for geographers. No. 310 C5. 1986.

Dai, F. C., C. F. Lee, and Y. Yip Ngai. "Landslide risk assessment and management: an overview." *Engineering geology* 64, no. 1 (2002): 65-87.

De Bruin, S., and A. Stein. "Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM)." *Geoderma* 83, no. 1-2 (1998): 17-33.

DESA, UN. "United Nations, Department of Economic and Social Affairs/Population Division (2016): The World's Cities in 2016 – Data Booklet (ST/ESA/ SER.A/392).

Devkota, Krishna Chandra, Amar Deep Regmi, Hamid Reza Pourghasemi, Kohki Yoshida, Biswajeet Pradhan, In Chang Ryu, Megh Raj Dhital, and Omar F. Althuwaynee. "Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya." *Natural hazards* 65, no. 1 (2013): 135-165.

Ercanoglu, Murat, and Candan Gokceoglu. "Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey)." *Engineering Geology* 75, no. 3-4 (2004): 229-250.

Gu, D., P. Gerland, F. Pelletier, and B. Cohen. "Risks of exposure and vulnerability to natural disasters at the city level: A global overview." *Population Division Technical Paper* 2015/2 (2015).

Guzzetti, Fausto, Alberto Carrara, Mauro Cardinali, and Paola Reichenbach. "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy." *Geomorphology* 31, no. 1 (1999): 181-216.

Kamp, Ulrich, Benjamin J. Growley, Ghazanfar A. Khattak, and Lewis A. Owen. "GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region." *Geomorphology* 101, no. 4 (2008): 631-642.

Lee, S. "Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data." *International Journal of Remote Sensing* 26, no. 7 (2005): 1477-1491.

Lee, Saro, and Touch Sambath. "Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models." *Environmental Geology* 50, no. 6 (2006): 847-855.

Nefeslioglu, H. A., C. Gokceoglu, and H. Sonmez. "An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps." *Engineering Geology* 97, no. 3-4 (2008): 171-191.

Pal, Nikhil R., and James C. Bezdek. "On cluster validity for the fuzzy c-means model." *IEEE Transactions on Fuzzy systems* 3, no. 3 (1995): 370-379.

USGS. "Landslides 101", USGS.gov, https://landslides.usgs.gov/learn/ls101.php (accessed January 5, 2018)

Wang, Liang-Jie, Kazuhide Sawada, and Shuji Moriguchi. "Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy." *Computers & Geosciences* 57 (2013): 81-92.

Wang, Rongxun. *An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic*. University of Wisconsin at Madison, 2008.

Wang, Weina, and Yunjie Zhang. "On fuzzy cluster validity indices." *Fuzzy sets and systems* 158, no. 19 (2007): 2095-2117.

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *IEEE Transactions on neural networks* 16, no. 3 (2005): 645-678.

Yalcin, A., S. Reis, A. C. Aydinoglu, and T. Yomralioglu. "A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey." *Catena* 85, no. 3 (2011): 274-287.

Yang, Lin, You Jiao, Sherif Fahmy, A. Zhu, Sheldon Hann, James E. Burt, and Feng Qi. "Updating conventional soil maps through digital soil mapping." *Soil Science Society of America Journal* 75, no. 3 (2011): 1044-1053.

Yilmaz, Işık. "Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat—Turkey)." *Computers & Geosciences* 35, no. 6 (2009): 1125-1138.

Yilmaz, Işık. "Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine." *Environmental Earth Sciences* 61, no. 4 (2010): 821-836.

Youssef, Ahmed Mohamed, Hamid Reza Pourghasemi, Zohre Sadat Pourtaghi, and Mohamed M. Al-Katheeri. "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, no. 5 (2016): 839-856.

Zhu, A-Xing, Rongxun Wang, Jianping Qiao, Cheng-Zhi Qin, Yongbo Chen, Jing Liu, Fei Du, Yang Lin, and Tongxin Zhu. "An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic." *Geomorphology* 214 (2014): 128-138.

Zhu, A-Xing. "A personal construct-based knowledge acquisition process for natural resource mapping." *International Journal of Geographical Information Science* 13, no. 2 (1999): 119-141.

Zhu, A-Xing, Lin Yang, Baolin Li, Chengzhi Qin, Tao Pei, and Baoyuan Liu. "Construction of membership functions for predictive soil mapping under fuzzy logic." *Geoderma* 155, no. 3-4 (2010): 164-174.

Zhu, A-Xing, and Lawrence E. Band. "A knowledge-based approach to data integration for soil mapping." Canadian Journal of Remote Sensing 20, no. 4 (1994): 408-418.

Zhu, A-Xing, B. Hudson, J. Burt, K. Lubich, and D. Simonson. "Soil mapping using GIS, expert knowledge, and fuzzy logic." *Soil Science Society of America Journal* 65, no. 5 (2001): 1463-1472.

Approved _____* on behalf of Prof. A Zhu

Advisor Title Professor

Dept. of Geography

Date 5/3/18

* Department Chair