8-2014

Predictive Model of Archaeological Sites of the Hopi Reservation of Northeastern Arizona

Jerry Thomas Holton Jr.
University of Redlands

University of Redlands

Predictive Model of Archaeological Sites of the Hopi Reservation of Northeastern Arizona

A Major Individual Project submitted in partial satisfaction of the requirements
for the degree of Master of Science in Geographic Information Systems

by

Tom Holton

Mark Kumler, Ph.D., Committee Chair
Douglas Flewelling, Ph.D.

August 2014

Predictive Model of Archaeological Sites of the Hopi Reservation of Northeastern Arizona

by

Jerry Thomas Holton, Jr.

The report of Jerry Thomas Holton, Jr. is approved.

Douglas Flewelling, Ph.D.

Mark Kumler, Ph.D., Committee Chair

August 2014

Acknowledgements

In final conclusion, I would like to thank God first and then my parents who have stood strong by my side since the beginning of the MS GIS program at the University of Redlands and my whole life. I would also most definitely like thank to Dr. Mark Kumler, my advisor, for his support, patience, and excellent technical advice throughout the year. I also want to thank my client, Dr. Wes Bernardini, for his flexibility, patience, and being an all-around excellent resource as I worked through my project. In addition, I would like to thank Michael Heilen, of Statistical Research, Inc., who gave me numerous tips on how to develop and improve the predictive models. I also want to thank the faculty of the MS GIS program, Dr. Fang Ren, Dr. Douglas Flewelling, Dr. Ruijin Ma, and Dr. Russell Weaver. All of them provided me with excellent technical and personal advice and assisted me while developing my project. Finally, I want to thank all my friends from Cohort 24, which is all of them. Without their friendship, support, knowledge of Geographic Information Systems science, and amazing weekend extravaganzas, surviving this year might not have been possible.

Abstract

Predictive Model of Archaeological Sites of the Hopi Reservation of Northeastern Arizona

by

Jerry Thomas Holton, Jr.

The Predictive Model for Archaeological Sites in the Hopi Reservation of Northeastern Arizona was developed to assist archaeologists in minimizing their study area for locating archaeological sites. Extensive research exists on predictive models for locating archaeological sites since the 1970s because many study areas are too large for archaeologists to cover on foot. The archaeological site types for these models were Habitation, Rock Art, and Scatter and were established between 500-1500 C.E. The independent variable categories for these models developed in ArcGIS 10.2 were based on topography, water resources, and vegetation. The logistic regression model in the Statistical Package for the Social Sciences (SPSS) was the selected statistical approach for building the predictive models. The final seven archaeological site type predictive surfaces were then created in the ArcGIS 10.2 Raster Calculator based on the coefficients created and statistically significant independent variables determined from the logistic regression models.

Table of Contents

Table of Figures

List of Tables

List of Acronyms and Definitions

DEM – Digital Elevation Model
CRM – Cultural Resource Management
GAP – Gap Analysis Program
GIS – Geographic Information Systems
NAD – North American Datum
PCA – Principal Component Analysis
SPSS – Statistical Package for the Social Sciences
SRI – Statistical Research, Inc.
SSURGO – Soil Survey Geographic Database
STATSGO – State Soil Geographic Database
TPI – Topographic Position Index
USGS – United States Geological Survey

Chapter 1 – Introduction

The Predictive Model of Archaeological Sites in the Hopi Reservation of Northeastern Arizona was designed to assist Dr. Wesley Bernardini, Associate Professor of the Department of Sociology and Anthropology at the University of Redlands, to better determine where to place survey transects to locate archaeological sites. The model created raster surfaces indicating the probability of finding three archaeological site types in the Hopi Reservation of Northeastern Arizona [Figure 1-1]. The site types were Habitation, Scatter, and Rock Art. The logistic regression model was the selected statistical approach that was executed in the Statistical Package for the Social Sciences (SPSS) to determine what independent variables and associated weights to use in creating the final predictive raster surfaces for the archaeological site types. In accomplishing this, archaeological survey efforts could be improved in the Hopi Reservation.

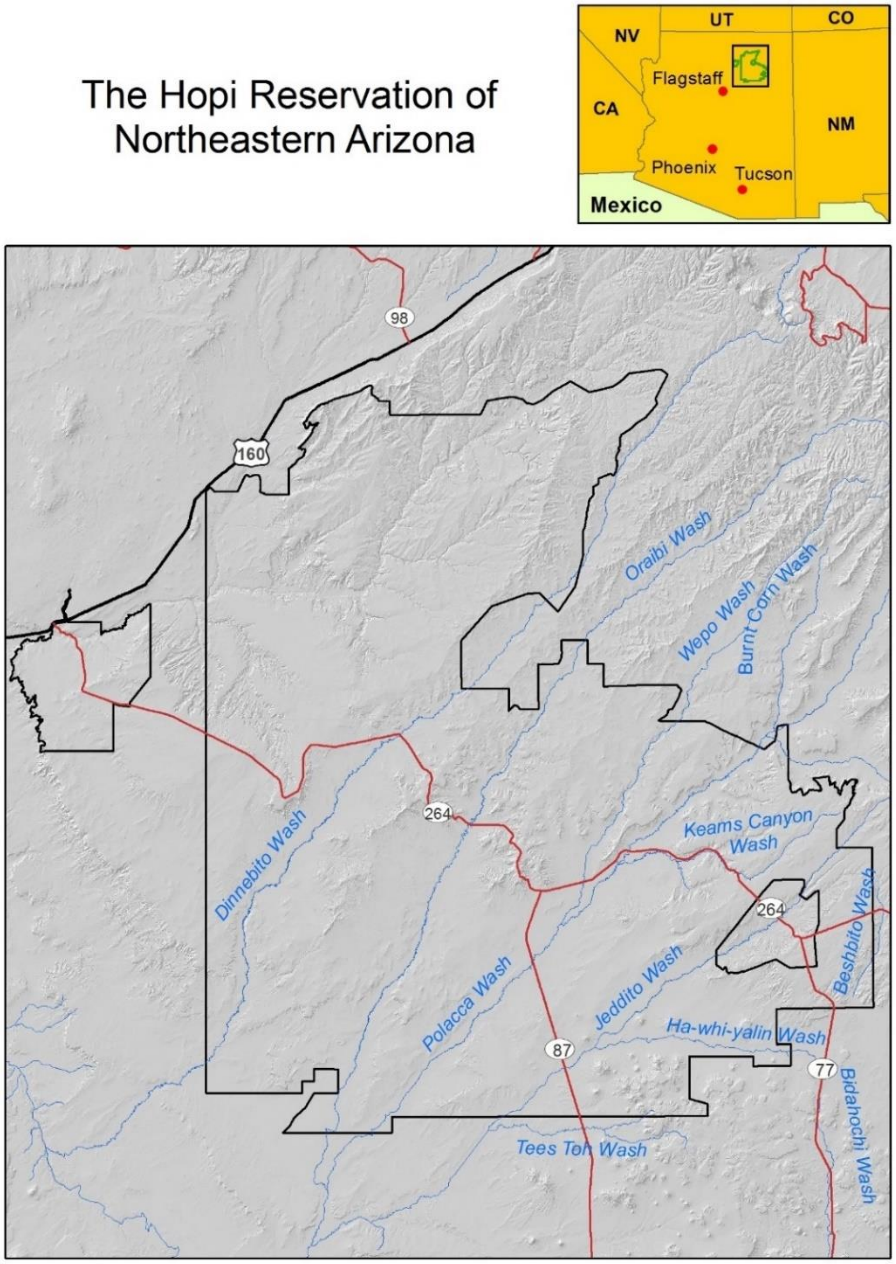


Figure 1-1: Reference map of the Hopi Reservation.

1.1 Client

Dr. Wesley Bernardini was the client for the project. He provided vector data of the current archaeological sites within the Hopi Reservation and its boundaries. He required a predictive model for archaeological sites in the Hopi Reservation to create a raster surface that indicates the likelihood of finding archaeological sites. Dr. Bernardini also required a file geodatabase to house the current data and that accommodates for future data, and to standardize projections for all data.

1.2 Problem Statement

Archaeologists are confronted with the problem of locating archaeological sites in extremely large study areas. This requires extensive field work and time in order to cover these study areas. Since the 1970s, developing predictive archaeological site models has become a common strategy among many archaeologists to increase the likelihood of locating sites (Kohler & Parker, 1986). As a result, Dr. Bernardini wanted to develop a predictive model for archaeological sites based on environmental variables (e.g., accessibility to water resources, aspect, slope, etc.). The outputs of this model were created to help focus his efforts and reduce the amount of fieldwork in the 1,619,936-acre Hopi Reservation (Bernardini, personal communication, September 30, 2013).

1.3 Proposed Solution

By developing a predictive model for archaeological sites in the Hopi Reservation, it was important to discuss the solution to the project. The strategy was influenced by a previous predictive archaeological model created by Statistical Research, Inc. (SRI), based out of Albuquerque, New Mexico.

The topographic, water resource, and vegetation variables for the model were created in ArcGIS 10.2. The logistic regression models were conducted in SPSS. Once the weights for each environmental variable were created based on the coefficients from the logistic regression model, these weights were then used in the weighted overlay analysis in ArcGIS 10.2. Outputs were created for the three archaeological site types. According to Fish (2013), The Learning Center of the Southwest, states that scatter sites are “entirely of artifacts and lacking associated features. Some artifact scatters may be comprised of a single material, such as a flaked stone or ceramics, whereas others encompass multiple artifact types.” Habitation sites range from ephemeral campsites to massive villages. Rock art sites are made up of pictographs and petroglyphs (Fish, 2013). As a result, raster surfaces indicating the likelihood of finding these three archaeological site types were created for the Hopi Reservation.

1.3.1 Goals and Objectives

This project contained one goal and five objectives. The overarching goal of the project was:

1. To create predictive models for three archaeological site types: Habitation, Rock Art, and Scatter.

In order to accomplish the project goal, five objectives were met:

1. Create a file geodatabase to store the appropriate data for building the models and other maps necessary for the project.

2. Conduct analyses for topographic, water resource, and vegetation independent variables to be included in the model.
3. Build logistic regression models in SPSS.
4. Create predictive raster surfaces for three archaeological site types.
5. Create a workflow for the client to model predictions when new data are added.

1.3.2 Scope

The scope of the project entailed creating a file geodatabase to house the appropriate GIS data to conduct the topographic, water resource, and vegetation independent variable analysis. The logistic regression models were then built in SPSS. The third component of the project was creating the predictive raster surface outputs in ArcGIS 10.2. Finally, a workflow was created that explained in detail how to update and run the model.

1.3.3 Methods

Once the file geodatabase was constructed, the topographic, water resource, and vegetation independent variables were created that were used as independent variables in the logistic regression model. The topographic variables included in the model were slope, north-south aspect, east-west aspect, local elevation change, terrain texture, shelter, topographic position index (TPI), and cost surface. The water resource variables utilized were cost to traverse to streams, cost to traverse to major streams, cost to traverse to springs and seeps, cost to traverse to water bodies, cost to traverse to all stream junction nodes, and cost to traverse to major stream junction nodes. Vegetation richness with a 100-meter and 500-meter radius were the only vegetation variables utilized in building models.

Various techniques were utilized to create the topographic variables. These variables were all created from a 10-meter United States Geological Survey (USGS) Digital Elevation Model (DEM) downloaded from http://nationalmap.gov/viewer.html. The seven topographic variables were then created in ArcGIS 10.2 utilizing various geoprocessing tools.

The water resource variables were the next environmental variables created. The stream flow lines, water bodies, and seeps and springs feature classes were all downloaded from the National Hydrography Dataset Plus website (http://www.horizon-systems.com/nhdplus/NHDPlusV1_CO.php). All water resource variables analyzed the relative difficulty of traveling across the landscape to a particular water resource (e.g., seeps and springs, major streams, water bodies). Major stream lines were identified by employing the Strahler Stream Order Methodology for determining the highest order streams within the reservation (Strahler, 1957). Stream orders 3-5 were then determined to be the highest order streams in the reservation, which allowed them to be designated as the Major Streams feature class. Major stream network nodes were also identified by utilizing various geoprocessing tools. The cost to traverse to major streamlines, all streamlines excluding the major stream lines, springs and seeps, water bodies, all stream junction nodes excluding the major stream junction nodes, and major stream junction nodes were all created in ArcGIS 10.2.

Vegetation data were downloaded from the Gap Analysis Program (GAP), a National Land Cover Data program of the USGS (http://gapanalysis.usgs.gov/

gaplandcover/). Vegetation was then examined by analyzing vegetation richness with a 100- and 500-meter radius within the reservation in ArcGIS 10.2.

Once the variables were created, the values for each environmental variable associated with each site type were exported out of ArcGIS 10.2 and were input in the logistic regression model. The dependent variables considered were the archaeological site types and the archaeological non-site types. Archaeological non-site types also had to be created for the model because the dependent variable in a logistic regression analysis is binary. The independent variables were topographic, water resources, and vegetation. Once the model was constructed and run, the coefficients generated by the model were used as weights for the statistically significant variables in the weighted raster overlay analysis. From this, the Raster Calculator was run for the three site types, which created various raster surfaces. There were multiple predictive surfaces created for each archaeological site type because the logistic regression analysis was run various ways to employ various strategies to create the most accurate and valid models.

1.3.4 Audience

The audience for this report is the client, Dr. Wesley Bernardini. Dr. Bernardini has a strong background in geographic information systems and sciences. Therefore, the final report is written in such a manner that he understands the majority of the technical concepts and language, although, there is a significant amount of clarification of technical terminology, concepts, and acronyms throughout the report due to the specialized nature of the project.

1.4 Overview of the Rest of this Report

Chapter 2 addresses the project background and literature review of research relevant to this project. Chapter 3 discusses systems analysis and design. Chapter 4 addresses the file overall database design for creating the models model. Chapter 5 discusses project implementation. Chapter 6 discusses the results and analysis of the report. Chapter 7 makes future recommendations as to where there could be improvements to the predictive model.

Chapter 2 – Background and Literature Review

It is common practice in the United States that federal land managers contract archaeologists to conduct surveys on federal lands to locate archaeological sites. Because of this, there has been development in the popularity of sensitivity models due to their ability to predict where settlement sites likely exist (Kohler & Parker, 1986). In the article *Predictive Models for Archaeological Resource Location* Timothy A. Kohler and Sandra C. Parker put forward two approaches to predictive modeling: the empiric correlative approach and the deductive approach. According to Kohler and Parker (1986), the empiric correlative model factors in environmental variables as being the main predictors of settlement location. This approach is designed to predict settlement areas that have similar environmental features, such as proximity to water resources, and generally does not consider social factors in the model. The deductive approach examines spatial behavior and can better answer questions as to why groups determine where they conduct their activities (Kohler & Parker, 1986). The predictive model created in this project followed an empirical approach because of the environmental variables that were utilized in the model (e.g., vegetation richness, local elevation change, shelter, and aspect) and did not consider social variables.

Heilen et al. (2012), states that archaeological sensitivity models have been used for Cultural Resource Management (CRM) since the 1970s. The Anthropological Studies Center of Sonoma State University states the following about CRM:

> …inventorying [cultural] sites, evaluating them, and at times mitigating the adverse effects of development projects and construction. Cultural resources are the remains and sites associated with human activities and include the following: prehistoric and ethnohistoric Native American archaeological sites; historic and archaeological sites; historic buildings; elements or areas of the natural landscape which have traditional cultural significance (Sonoma State University, 2008).

Because of this need, archaeological predictive models have been developed to assist in locating these sites in large study areas.

This chapter addresses the models' main components that had to be researched in order for successful model development. These areas were the model variables, the logistic regression statistical approach to model development, applications of the logistic regression model in predictive modeling, and approaches to model validation and accuracy. This chapter also discusses the principal component analysis (PCA) that is another method that can be used in building a predictive model, but was ultimately not used in this project. The chapter concludes by providing a high-level understanding of how each of these topic areas were implemented in building models.

2.1 Model Variables

There are many approaches that have been used for determining which independent variables should be utilized in predictive modeling. Statistical Research, Inc. (SRI) staff developed a predictive archaeological site model in southern New Mexico in 2012 and created variables within five categories:

- Water resources
- Soils
- Topography
- Historical-period resources
- Vegetation

Many of these variables were implemented in this predictive model for archaeological sites in the Hopi Reservation.

The topographic variables utilized for SRI's model were all created from a 10-meter DEM of the state of New Mexico. These variables were slope, north-south aspect, east-west aspect, local elevation change, terrain texture, shelter, elevation above water, and cost surface. Slope is a very common variable utilized in predictive modeling because the slope of the terrain can have a strong impact on archaeological site locations. Aspect is also very important because a site's orientation can have direct impact as to the level of exposure to sun and wind (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). According to Heilen et al. (2012), terrain texture is a measure of terrain roughness. Shelter was also an important variable for modeling for site location. This is "the degree to which topographic features offer shelter from the weather, sun, or even visibility" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Local elevation change is important for modeling site locations because a rough terrain can "inhibit day-to-day activities and travel to and from sites" (Kvamme, 1988). The cost surface variable was created to measure the relative cost of moving across the landscape. This variable was useful in creating variables that measure the distance between a raster cell and the closest resource of a given type (e.g., water bodies) (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). These variables were used in creating predictive archaeological site models for southern New Mexico.

The model developed by SRI staff also utilized soil variables that were extracted from the Soil Survey Geographic Database (SSURGO) and State Soil Geographic Database (STATSGO). SRI staff used multiple soil-attribute variables in the model because they found that for many sites each modeling unit was located along the edges of soil map units (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Some of these variable were available water capacity, pH level, organic matter, surface-horizon thickness, and sodium absorption ratio. SRI staff also created additional soil variables such as cost to traverse to soil-texture boundary, standard deviation in soil-texture boundary index, and range in organic matter. Cost to traverse to soil-texture boundary examined the cost to traverse to a soil mapping unit boundary. The standard deviation in soil-texture index variable was created to examine the variation in soil texture between adjacent soil-mapping units. The range in organic-matter content was created because it was observed that sites close to soil-mapping unit boundaries were "located on units that has substantially different organic matter contents than immediately adjacent soil-mapping units" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012).

Water resource variables were also created to examine their availability on the landscape. The variables taken into consideration for water resources were cost to traverse to streamlines, cumulative drainage of nearest stream segment, cost to traverse to major streamlines, cost to traverse to springs or seeps, cost to traverse to water bodies, cost to traverse to stream-network nodes, cost to traverse to major stream-network nodes, and elevation relative to water. All these variables measured the relative difficulty of traversing the landscape to reach the specified water resource (e.g., cost to traverse to springs or seeps). Elevation relative to water was another water resource variable that SRI staff utilized in building models. Because these models were developed in the canyon and mesa land of southern New Mexico, it was important to examine not just the horizontal cost to traverse to a water resource but also to examine the vertical distance, for example, an area might horizontally be in close proximity to water but is separated from it by a cliff. For this reason, elevation relative to water was included in the building models (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012).

Another variable taken into consideration by SRI staff was historical period resources. Heilen et al. states,

> Historical-period sites are often located near transportation routes, such as trails, wagon roads, and railroads because much historical-period land use in the US West was dependent on the use of transportation for exploration and population migration as well as the redistribution of goods and materials. Therefore, proximity to such features of the built environment is often considered a primary determinant of historical-period settlement behavior (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012).

Similar to the water resource variables, all the names of the historical-period variables are self-explanatory in how they influenced site establishment (e.g., cost to traverse to historical-period places). These variables included cost to traverse to historical-period transportation routes, cost to traverse to historical-period places, and cost to traverse to [water] tanks.

Vegetation was the last variable that was considered in the variable analysis. SRI staff utilized this variable because they believed there were strong associations between site location and vegetation type. SRI staff only examined two variables, vegetation type and richness. SRI staff considered vegetation richness because they believed sites that were “located near the edges of multiple vegetation types have greater access to a wider variety of resources than sites located in an area with uniform vegetation” (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012).

The article *GIS Reveals Basis for Ancient Settlement Location*, by Dr. Terrance L. Winemiller, Associate Professor of Archaeology, Auburn University of Montgomery, Alabama, discusses a similar predictive model development project that was created for covering most of northwestern Yucatán, Mexico (Winemiller, 2014). The underlying theory behind this model was that the Mayas established settlements close to water sources, especially small cenotes (natural wells). There was additional analysis of other environmental variables, such as rainfall, climate, and soil type, but no significant relationships were found between site type and these variables (Winemiller, 2014). Essentially, what was ascertained from this research is that water resource variables were the strongest site determinant in the study. This could be due to the fact that the terrain was relatively flat, which caused site location to not be dependent on topography.

2.2 The Principal Component Analysis

The principal components analysis (PCA) has been used in predictive models to reduce the number of independent variables that were originally created for building models for simplification and also to eliminate multicollinearity among variables. Multicollinearity arises where there are very strong correlations among independent variables and the independent variables have no significant impact on explaining the dependent variable (Fattah, 2010). According to the Esri ArcGIS Resources Center website, the PCA tool found in the Spatial Analyst toolbox of ArcGIS 10.2 transforms the original raster data "in the input bands to a new multivariate attribute space whose axes are rotated with respect to the original space. The axes (attributes) in the new space are uncorrelated" (Esri, 2011). The purpose of this statistical technique is to eliminate multicollinearity or redundant variables. For example, elevation, slope, and aspect are derived from a DEM, from which most of the variance can be explained. The tool then creates a multiband raster that contains the same amount of bands as the specified number of components. The first principal component raster band contains the greatest amount of variance, the second contains the second greatest variance, and so on (Esri, 2011). Typically, principal component rasters containing 95 % of the variance will be utilized in building models. In order for the Principal Components Analysis tool to be run, the user must enter the following inputs: the input bands (environmental variables), the number of principal components to transform the original input bands, and the name of the statistics text file (Esri, 2011).

In order to shift and rotate the axes, the data are plotted onto a scatter plot. An ellipse then bounds the data points in the scatter plot [Figure 2.1] (Esri, 2011).

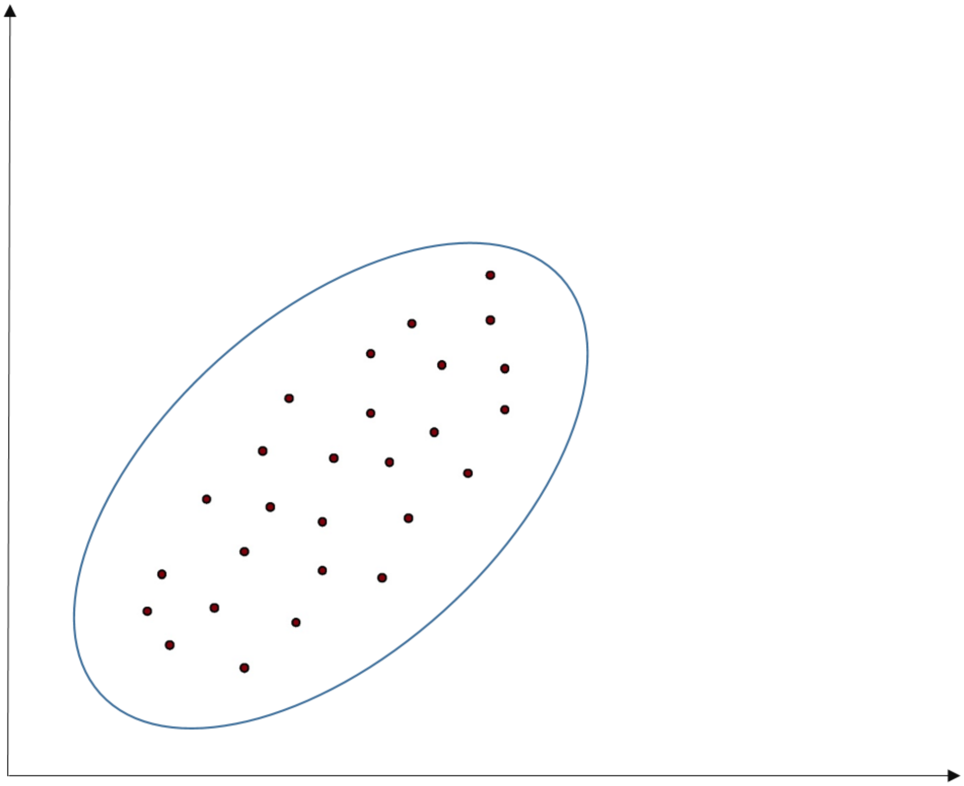


Figure 2-1: Data points bounded by a fitted ellipse. Adapted from ArcGIS 10.2 Help, How Principal Component Works (Esri, 2011).

Next, the software program determines the major axis of the ellipse. This component exhibits the most variation "because it is the largest transect that can be drawn through the ellipse. The direction of PC1 is the eigenvector, and its magnitude is the eigenvalue [Figure 2-2]. The angle of the x-axis to PC1 is the angle of rotation that is used in the transformation" (Esri, 2011).

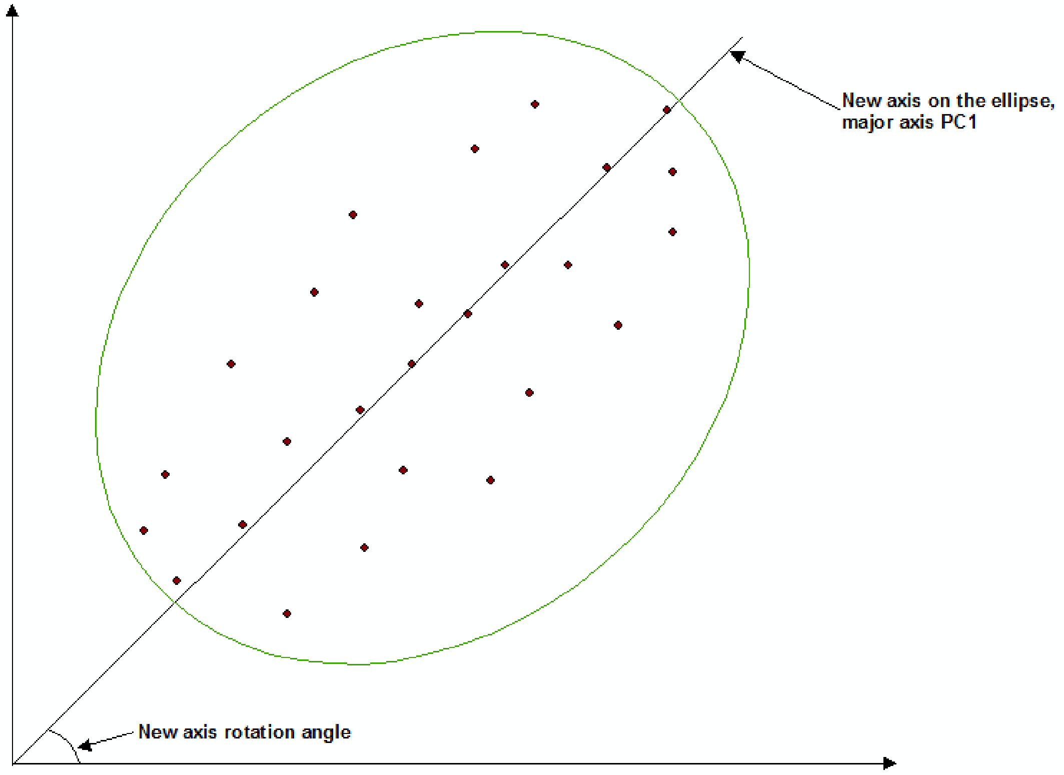


Figure 2-2: First Principal Component. Adapted from ArcGIS 10.2 Help, How Principal Component Works (Esri, 2011).

The orthogonal line that is directly perpendicular to the first principal component is then calculated (Esri, 2011). This line created is now the second principal component and is also the new axis of the original Y-axis [Figure 2-3]. The axis now describes the greatest variance that was not described by the first principal component (Esri, 2011).

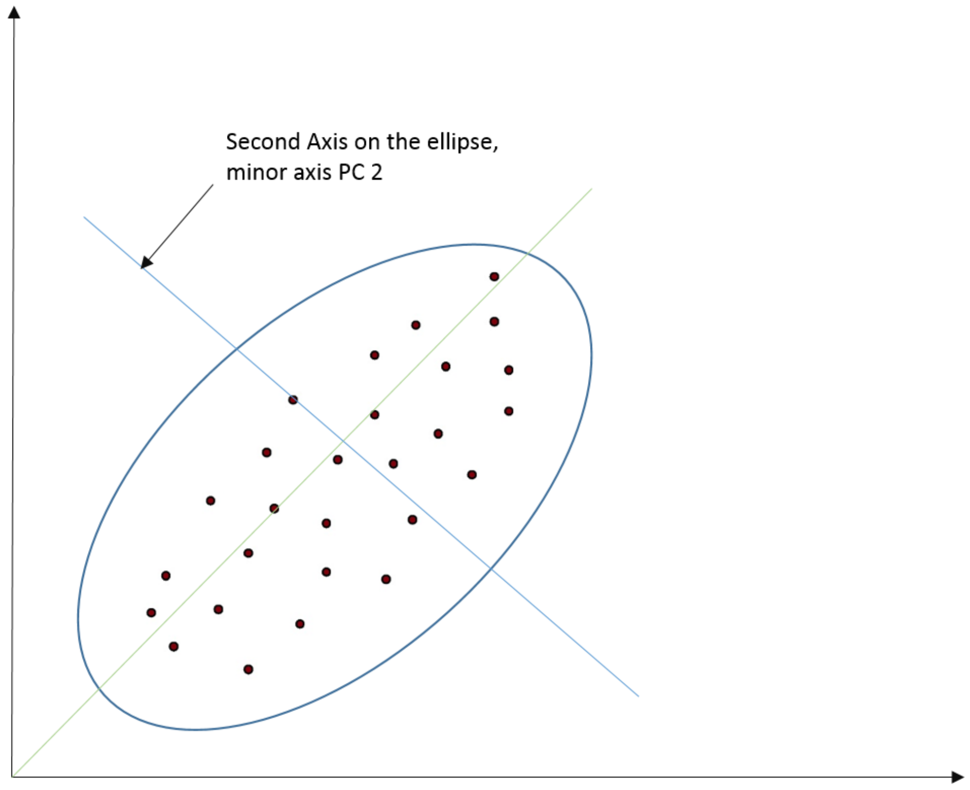


Figure 2-3: Second principal component. Adapted from ArcGIS 10.2 Help, How Principal Component Works (Esri, 2011)

A line formula is then created that defines the shift and rotation previously mentioned by utilizing the eigenvectors, eigenvalues, and the covariance matrix that was created by the input multiband raster (Esri, 2011).

It is important to understand how the percent variance is determined. The percent variance identifies how much variance each eigenvalue captures. This is used to interpret the PCA results. The formula to determine the percent variance is the following:

Percent Variance = (eigenvalue * 100)/Sum

It is recommended to use the principal components whose eigenvalues contain the majority of the variance to build models (Esri, 2012).

2.3 Logistic Regression Approach in Predictive Modeling

The logistic regression analysis is the statistical approach that was selected for this project. Logistic regression is a multivariate regression specification for a dichotomous dependent variable and independent variables that are either categorical or continuous (Field, 2005). In essence, categorical dichotomous variables are binary. For this model, the categorical variables were archaeological site presence or absence. The purpose of the logistic regression model was to determine areas in the Hopi Reservation where there is likelihood of archaeological site presence or absence, based on multiple independent variables. In a linear regression equation where there are multiple independent variables, the relationship between the dependent variable and a given independent variable is captured by an estimated coefficient on that variable. For example,

$$Y_i = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \varepsilon_1$$

where b_0 is the constant, ε_1 is residual term, b_n is the regression coefficient corresponding to the independent variable, $X_{n,}$ (Field, 2005). In a logistic regression model, the objective is to predict the probability of *Y* occurring given the known independent variable values.

In a logistic regression, where there are multiple independent variables, the equation creates a probability range between 0 and 1. This indicates that a value close to 1 suggests that *Y* is likely to occur and a value close to 0 suggests that *Y* is not likely to occur (Field, 2005). This is expressed in the following equation:

$$\mathrm{P(Y)} = 1 + \frac{1}{1+(\mathrm{e}^\wedge b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \varepsilon_1)}$$

where P(Y) is the probability of *Y* occurring, e is the base logarithm, b_0 is the constant, and the other coefficients listed are the same as the previous logistic regression analysis discussed above. When the logistic regression model is run, the value of the coefficients needs to be estimated so the equation can be solved (Field, 2005). The objective is that the chosen model will be one that when the independent variable values are placed in the model, the results of the *Y* values will be closest to the observed values—the archaeological sites/non-sites that have already been located (Field, 2005). The primary goal of the logistic regression model is to fit a model to the observed data that allows the researcher to estimate values of the dependent variable from the known values of the independent variables (Field, 2005).

Field then begins to discuss what method to use to determine the most significant variables for determining the final model, whether it be stepwise or the forced-entry method. He states that most people believe the stepwise method is most appropriate for exploratory work, therefore, it is most suitable for the predictive model, considering no prior research has been done on this topic in the Hopi Reservation.

In the forward stepwise method, the model is initially defined only by the constant, b_0. This method is then able to detect the independent variable that best predicts the dependent variable, *Y.* It selects the independent variable based on the variable that has the highest correlation with the dependent variable. The forward stepwise method then selects the next variable that has next highest correlation, or semi-partial correlation, with the independent variable. Essentially, this methodology is able to select the independent variable that can explain most of the dependent variable. Semi-partial correlation can be explained by how the software program correlates each of the independent variables with the dependent variables “while controlling for the effect of the first predictor [independent variable]” that has already been selected (Field, 2005).

The backward step-wise methodology begins by including all the independent variables in the model and then determines the level of contribution of each one by examining the significance value of the t-test of each independent variable (Field, 2005). When utilizing the stepwise method, Field recommends using the backward methodology because of suppressor effects. Suppressor effects influence when an independent variable has significant effects, but only when another variable is held constant. According to Field, the forward step-wise is more likely to not include independent variables that are impacted by suppressor effects (Field, 2005). However, in this project the two methodologies produced the same results.

These concepts and principles which have been previously discussed comprise the background of the logistic regression model. This is important to comprehend in order to effectively create a predictive archaeological site model utilizing logistic regression. This will also allow one to understand how logistic regression has been applied in other predictive modeling applications.

2.4 Applications of the Logistic Regression Model in Predictive Modeling

In the article *Predictive Mapping of Blackberry in the Condamine Catchment Using Logistic Regression and Spatial Analysis*, the logistic regression modeling approach was utilized for predictive mapping (Apan, Wells, Reardon-Smith, McDougall, & Basnet, 2008). In this article, the authors posed that a logistic regression model is “used to predict the probability of occurrence of an event as a function of the independent variables. It is useful when the observed outcome is restricted to two values, which usually represent the occurrence or non-occurrence of an outcome event” (Apan et al., 2008). This statistical approach does not adopt the approach that the relationship between the independent and dependent variables is linear, meaning a 1:1 correlation. It also does not assume that the dependent variable is normally distributed (Apan et al., 2008). This is important for archaeological predictive modeling because archaeological site locations would probably never be normally distributed (Field, 2005), meaning they tend to be clustered due to the environmental conditions that dictated their location.

The authors selected 2,277 randomly selected sample points that contained either weed or no weed observations. From this, they chose 1,592 points for the training set of the analysis. The authors utilized the forward likelihood ratio [forward like-wise] method, which allowed them to determine the number of categorical variables to be used in the model, as opposed to the SRI staff who used the PCA approach to limit the number of variables to be used in their model. After each of the variables had been tested with “an assessment of their related statistics, a model was finally selected. The following equation was implemented using the ArcGIS Spatial Analyst raster calculator to produce the predictive map” (Apan et al., 2008):

$$\textit{Pred\_bberry} = 1 \text{ div } (1 + \exp(-(-7.751 + (-.002*[\textit{dist\_stream}]) + (.000*[\textit{dist\_border}]) +$$

$$(-.037 * [\textit{fpc}]) + (.015 * [\textit{elev}])))))$$

By using the forward-likelihood ratio [forward like-wise], the authors were able to determine the most important variables to include in the final model. The final output map was a predictive raster surface where each cell value ranged from 0 to 1, where 1 indicated the highest likelihood of finding blackberry (Apan et al., 2008).

In the article *Predictive Models of Archaeological Site Distributions in New Zealand* J.R. Leathwick also used the logistic regression modeling approach for predicting archaeological sites with ArcGIS software. Leathwick used the Generalized Additive Model to fit the logistic regression model to the data in order to estimate the independent variables. Leathwick asserts that this approach is advantageous "in that the relationship between the response and the continuous predictor variables are defined from the data using scatter plot smoothers, rather than more inflexible parametric terms used traditionally for such analysis" (Leathwick, 2000). The regression was then fitted using the backwards stepwise procedure, where all variables were initially used and then dropped once their significance using the T-test was determined. Leathwick stated that the advantage of the logistic regression approach, as stated earlier, is that it has the ability to handle non-normally distributed data, such as the presence/absence archaeological site data used in this study (Leathwick, 2000). The idea behind this is that the data points in the scatter plot can be smoothed "by fitting a line to the data" (NetMBA, 2010). Leathwick assessed the individual contributions of each independent variable by "using the residual deviance when dropping each [variable]" (Leathwick, 2000). The predictive archaeological site map was then produced using the same environmental dataset points on a 1-km grid (Leathwick, 2000).

2.5 Accuracy and Validation

It is recommended to validate and analyze the accuracy of the results of models. In the model developed by Apan et al. (2008) validation and accuracy measures were conducted to analyze these results. This was necessary to determine whether the outputs were valid and could be used for decision-making. The authors used the Hosmer-Lemeshow chi-square test of goodness-of-fit to examine the overall fit of the model. The desired outcome of this test is to have a finding of non-significance, e.g., p-value above .05 significance level (SPSS). This indicates the model adequately fits the data. The authors also suggested the omnibus test of model coefficients to test "whether the model with the predictors is significantly different from the model with only the intercept" (Apan et al., 2008). The desired outcome of this test is to have a finding of significance e.g., p-value below .05 significance level. This also indicates that the data adequately fits the model. The authors state, "This means that at least one of the predictors is significantly related to the response variable" (Apan et al., 2008). In the same study, the authors used classification tables to assess the accuracy of the model. The idea behind this test is to determine what percent of the model is correctly classified. According to the authors, the goal is to have a perfect model that scores a 100 % correct. These tests were conducted to examine whether the variables selected for building models were valid and could adequately explain the dependent variable, site/non-site locations (Apan et al., 2008).

2.6 Summary

In the course of examining much of the current literature on the topic of predictive models, an overall strategy was formulated. The study conducted by the SRI staff had the greatest amount of influence on what independent variables were created to build models for the project. The work of Field (2005) also was very influential in how the logistic regression model was built in SPSS. Because of this work, the forward step-wise and forced-entry methods were selected to determine which independent variables were used for building the final models for each archaeological site type. The Hosmer-Lemeshow test was applied to examine the models' goodness-of-fit and the omnibus tests of model coefficients was utilized to test whether the independent variables were better able to predict the independent variables "better than by chance alone" (Denis, 2010). In order for a predictive surface to be created in ArcGIS 10.2, the algorithm suggested by Apan et al. (2008) was used in the Raster Calculator geoprocessing tool, with the coefficients created in SPSS to standardize the outputs of the predictive surface to create range from 0 to 1.

Once the model results were created, the models were assessed for accuracy. The percent correctly classified statistic was calculated to test the accuracy of the output, which was also utilized by Apan et al. (2008). As a result, this test provided a level of understanding of how accurate the final output maps were and where to focus archaeology study efforts in the Hopi Reservation.

Chapter 3 – Systems Analysis and Design

Because the client requested a predictive model for archaeological sites, it is necessary to clearly discuss the overall systems analysis and design of the project. This entails revisiting the problem statement in section 3.1 to clarify what was actually needed by the client. After discussing the problem statement, the project functional and non-functional requirements are then discussed in section 3.2. The project system design is then addressed in more detail in section 3.3, followed by the project plan in section 3.4.

3.1 Problem Statement

Archaeologists are confronted with the problem of locating archaeological sites in extremely large study areas. This requires extensive fieldwork and time in order to cover these study areas. Since the 1970s, developing predictive archaeological site models has become a common strategy among many archaeologists to increase the likelihood of locating sites (Kohler & Parker, 1986). As a result, Dr. Bernardini wanted to develop a predictive model for archaeological sites based on the following environmental variables: water resources, topography, and vegetation. The outputs of this model were created to help to focus his efforts and reduce the amount of fieldwork in the 1.6 million-acre Hopi Reservation (Bernardini, personal communication, September 30, 2013).

3.2 Requirements Analysis

The client had three functional requirements for the project. In order to create the predictive model, the data had to be appropriately housed and organized in a file geodatabase. These feature class data were archaeological site/non-site feature classes, hydrography feature classes, site and non-site values to points, convex hulls, and additional feature classes for reference maps (e.g., roads, political boundaries, and streams). The raster topographic, water resource, and vegetation independent raster variables for the model were housed outside of the file geodatabase in separate folders because of their large quantity. Because the predictive model was never a Python or ArcGIS ModelBuilder tool that could be run to create predictive archaeological site raster outputs, the predictive model requirements were only predictive raster surface maps indicating the likelihood of locating each of the three archaeological site types. Because it was not possible to create an automated tool, the third functional requirement was a workflow explaining how to update the model when new archaeological sites were needed to be added to increase its accuracy (Table 1).

The only non-functional requirement made by the client was that the model be tested for accuracy. The client also visually inspected the final outputs of the model to see whether the results were logical. The model was also tested utilizing the classification tables created in SPSS for model accuracy assessment.

Table 1. Functional and Non-functional requirements.

Functional Requirements	**Description**
File geodatabase	File geodatabase that houses feature datasets of archaeological site and non-site point feature classes, hydrography feature classes, feature classes for reference maps (e.g., roads, political boundaries), convex hull polygons for non-site random sampling, and points to value feature class that holds variable values associated with site and non-site points
Predictive model	Predictive model utilizing topographic, water resource, and vegetation variables to detect archaeological site type locations according to archaeological period
Predictive model workflow for updating model	Training manual explaining how to update predictive model when new archaeological sites are located
Non-Functional Requirements	**Description**
Model Assessment & Validation	The client visually assessed the model outputs to determine whether the results are acceptable. The model was tested and validated using the Percent Correctly Classified methodology

3.3 System Design

The file geodatabase was the first major component of the overall system design. This was developed not only for model development and reference maps, but also for the client's future use if he decides to update the model and for other GIS purposes. Once the file geodatabase was developed and organized, the overall system design was created [Figure 3-1]. The overall system design comprised the independent and dependent variables created in ArcGIS 10.2. These dependent variables—the archaeological site/non-site points—were used to extract the values of the environmental variables to those points and then were imported into SPSS in order to build models. Once the coefficients were generated for each site type by running the logistic regression analysis, these coefficients were then input to a formula in the Raster Calculator geoprocessing tool to create the predictive raster surfaces [Figure 3-1].

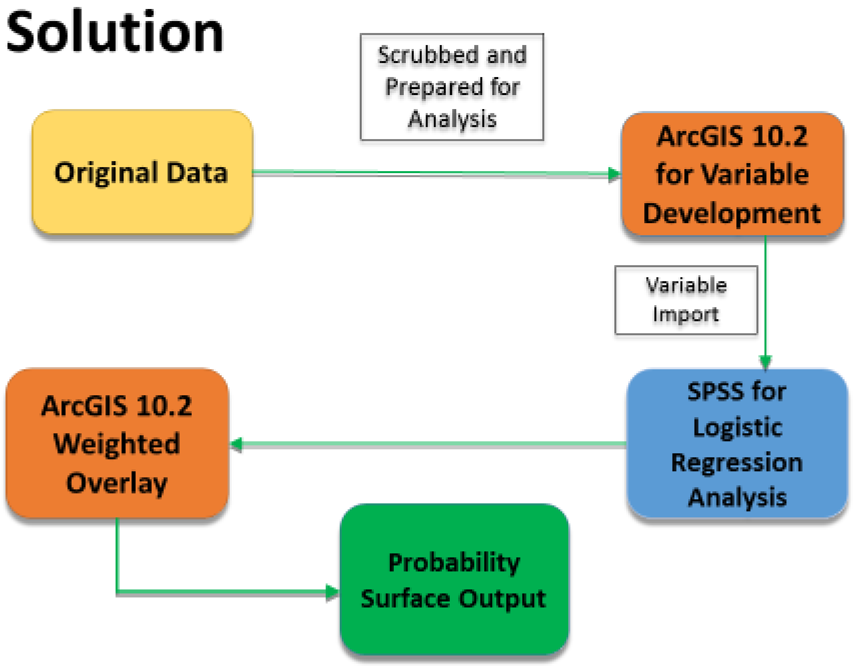

Solution
Original Data
Scrubbed and Prepared for Analysis
ArcGIS 10.2 for Variable Development
Variable Import
SPSS for Logistic Regression Analysis
ArcGIS 10.2 Weighted Overlay
Probability Surface Output

Figure 3-1: System design of predictive model workflow

3.4 Project Plan

The original project plan was divided into numerous milestones [Figure 3-2]. The geodatabase construction milestone of the project was divided into data scrubbing, data reprojection, and data acquisition and was projected to be completed by the end of February 2014. The model construction milestone was divided into the principal components analysis and cost to traverse analysis and was projected to be completed by May 15, 2014. Model assessment and testing were also divided into many tasks: comparing current archaeological sites to predicted locations, utilizing ArcGIS Data Reviewer, and other testing and validation measures. This milestone was projected to be completed by June 6, 2014. The original project plan also included a web application. This was to be completed by July 5, 2014. Project requirement revisions were projected to be completed by June 30, 2014. The final project report was projected to be completed by July 31, 2014 [Figure 3-2].

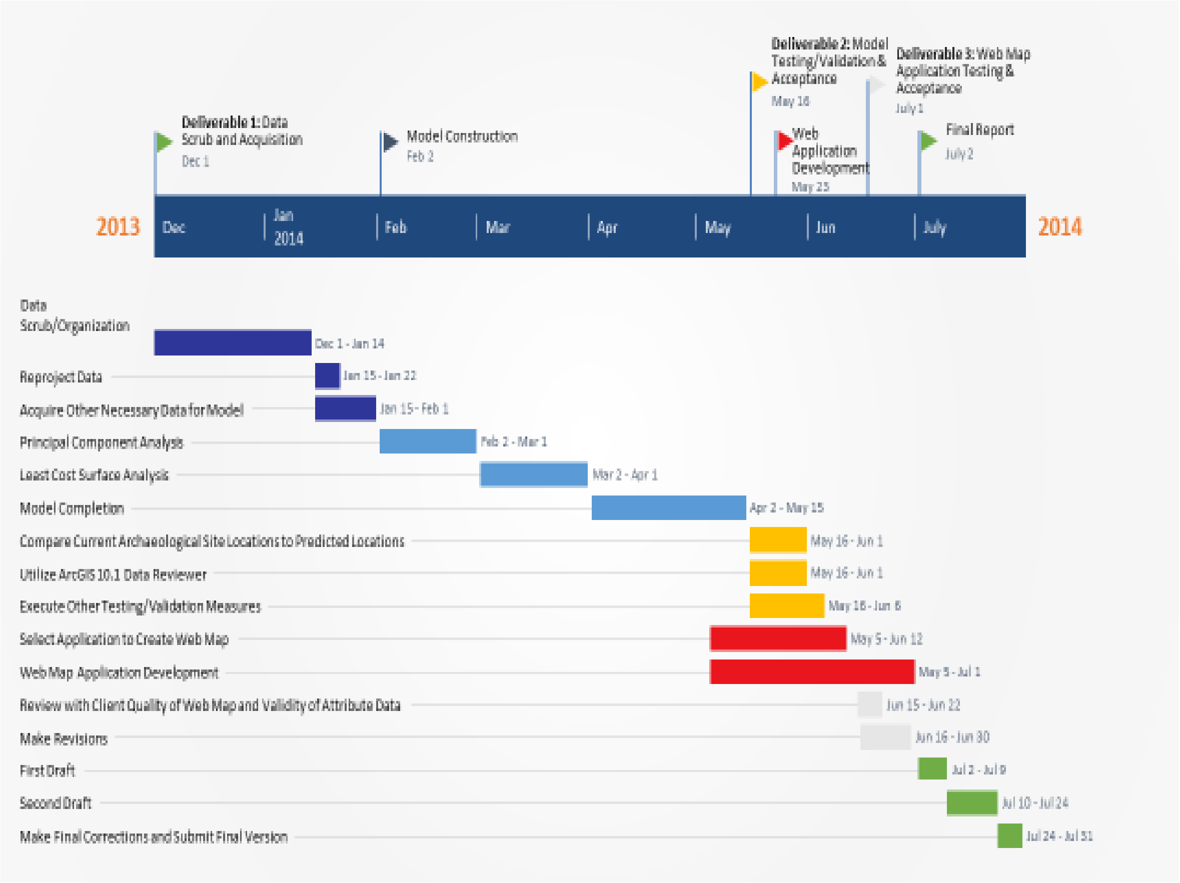


Figure 3-2: Original project plan

In the actual execution of the project plan, a large amount of time was allocated to geodatabase design and creation, and data collection. The completion of this milestone not only entailed the actual technical aspects of the file geodatabase scrubbing, design, and creation, but also required extensive planning and clarification of file geodatabase requirements with the client. This was expected to be completed by the end of February 2014. This milestone of the project was not actually reached until March 7 [Figure 3-2].

Milestones of the project such as the Principal Component Analysis and Least Cost Surface Analysis were projected to be completed by the end of April [Figure 3-1]. In the updated project plan, these milestones were eliminated and a new milestone was created, called simply Model Construction. The model was projected to be completed by May 15, 2014 [Figure 3-3]. This milestone was actually completed June 11, 2014. There were some technical difficulties during the model construction milestone of the project. The trigger point date, the date to change to the contingency plan, for model construction technical difficulties was May 15. The contingency plan for this was to seek outside technical assistance from experts in the field of spatial statistics and logistic regression analysis.

The project scope was also reduced. Originally, the client requested that the Hopi reservation be divided into three modeling units—Great Basin Conifer Woodlands, Great Basin Shrub Grassland, and Great Basin Desert Scrub, which would increase the number of predictive model map outputs significantly. Due to time constraints, the client agreed to the contingency plan to only require the model outputs be created for the whole Hopi Reservation rather than individual modeling units. This then only required model outputs for each site type, which were not based on archaeological period. The web application was also eliminated from the original project scope because it was determined unnecessary for the client's needs [Figure 3-3].

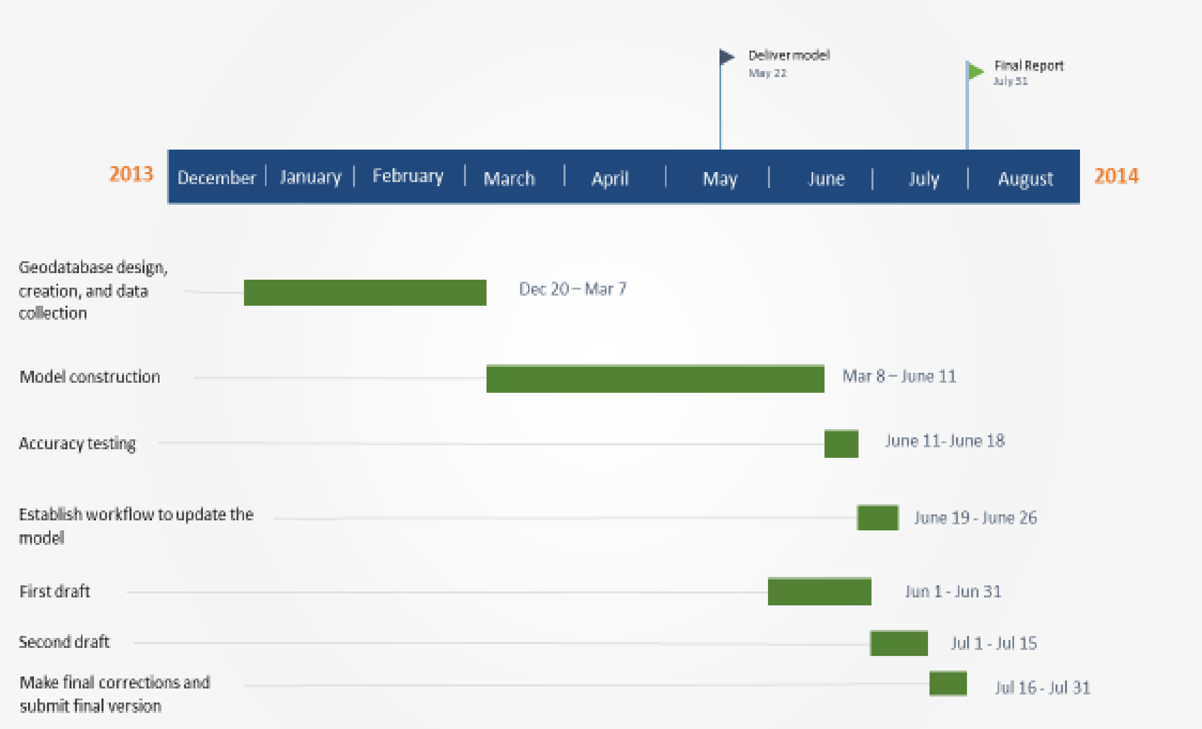


Figure 3-3: Revised project plan

3.5 Summary

In summary, the project plan and timeline changed significantly throughout the course of the project. Project scope was reduced significantly by eliminating the web application and also by creating models for the entire Hopi Reservation at once and not by individual model units, which would have tripled the amount of model outputs. Geodatabase design and model construction also took significantly longer than expected. With this said, it was important that there were the necessary trigger point dates set for when problems arose and also contingency plans. It was also necessary that many of the project milestones were completed in advance in order to have time to complete more difficult milestones.

Chapter 4 – Database Design

The overall database design was one of the most crucial components of the project, not only for model building purposes but also for the client's future use. Section 4.1 describes the conceptual data model and the underlying theory of how and why the model was developed. Section 4.2 discusses the logical model, how the appropriate data were organized within the file geodatabase, and how the independent raster variables were stored. Section 4.3 discusses the data sources of the project and provides justification for why these datasets were used. Section 4.4 addresses outside data collection efforts. Section 4.5 discusses the data scrubbing and loading component of the project using data provided by the client. Section 4.6 summarizes the chapter and emphasizes some of its more salient points.

4.1 Conceptual Data Model

The conceptual data model of the predictive model for archaeological sites of the Hopi Reservation of northeastern Arizona represented the relationships between the independent environmental variables and the final model outputs [Figure 4-1]. The independent variables were represented by three categories: topography, water resources, and vegetation. The conceptual data model factor addressed why each variable was used in creating the predictive model. For example, the topographic variables dictated the amount of protection or shelter that was provided, which played a role in site location. Vegetation was an indicator of biodiversity, which provided food resources. Water resources were important in that they provided a water resource for survival and travel. The final model output was a predictive surface indicating the likelihood of finding an archaeological site within the reservation [Figure 4-1]. The topographic and water resource independent variables were created from 10-meter USGS DEMs, while the vegetation independent variable was created from GAP vegetation data, which is also depicted in the model. The water resource feature classes required for the creation of the water resource independent variables were created from the National Hydrologic Dataset Plus (U.S. EPA, n.d.) [Figure 4-1].

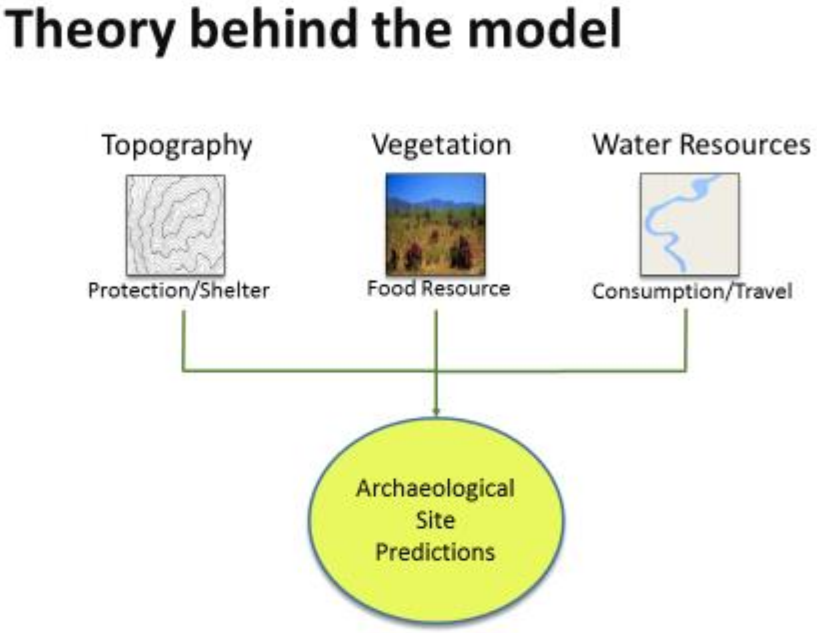


Figure 4-1: Conceptual data model

4.2 Logical Model

The majority of the data was stored in a file geodatabase in six feature datasets for the logical data model [Figure 4-2]. These feature datasets were Archaeological Sites, Hydrography, Convex Hulls, Surveys, Mapmaking Resources, and Values to Points. Within the archaeological site types feature dataset, there were multiple feature classes. These were Habitation, Scatter, Rock Art, and the Non-Sites. The Hydrography feature dataset contained the Water Bodies, All Streams, Major Streams, and Seeps and Springs feature classes. The surveys feature dataset contained only the All Surveys feature class. The Mapmaking Resources feature dataset contained the following feature classes for mapmaking: Major Cities, Roads, Reservation, Stream Washes, and US States. The Convex Hull feature dataset contained the convex hulls that were used as the constraining boundaries when the random non-site points were created for each archaeological site type [Figure 4-2].

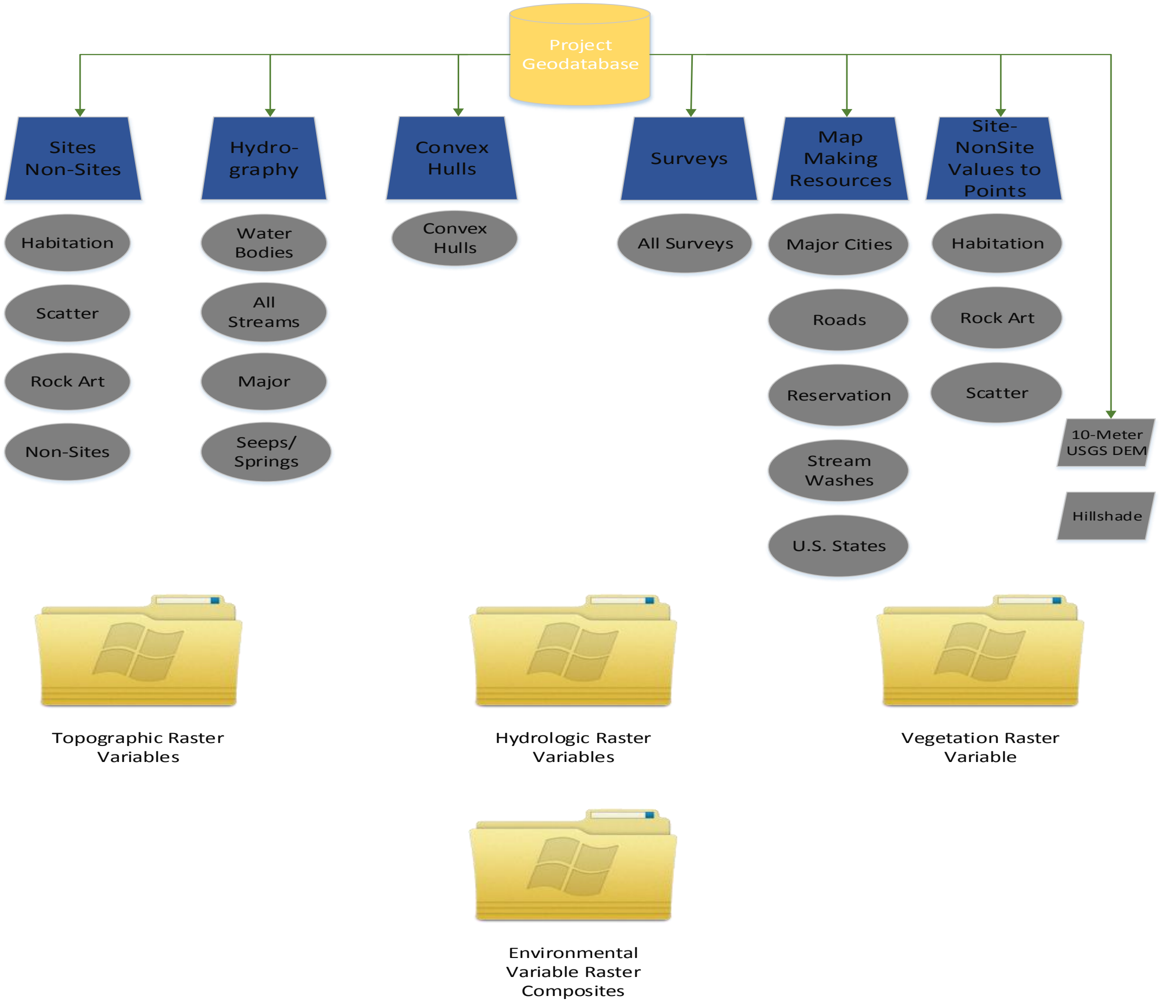


Figure 4-2: The logical data model.

The independent model raster variables were contained within folders outside the project geodatabase [Figure 4-2]. These variables were topographic raster variables, water resource raster variables, and the vegetation raster variable. The topographic raster variables folder contained eight independent model variables: slope, local elevation change, cost surface, shelter, terrain texture, east-west aspect, north-south aspect, and TPI. The water resource variables contained six independent model variables: cost to traverse to streams, cost to traverse to major streams, cost to traverse to seeps and springs, cost to traverse to water bodies, cost to traverse to stream nodes, and cost to traverse to major stream nodes. The vegetation raster variables contained two independent model variables: vegetation richness with 200- and 500-meter radii. All 16 independent variables were converted to a single composite variable, which was stored separately in the environmental variable raster composites folder [Figure 4-2].

4.3 Data Sources

Archaeological site point data were provided directly by the client that had been collected from his personal surveys as well as those of other associates. This data came in Excel and Esri shapefile formats. The Hopi Reservation boundary was also provided directly by the client in Esri shapefile format.

All hydrography data were downloaded from the National Hydrography Dataset Plus website of the USGS. The hydrography data was from Region 15, version 01. The flow lines, seeps and springs, and water bodies were the only datasets used from the NHD Plus website. The NHD Plus website is the official United States geospatial hydrologic framework dataset that was created by the US Environmental Protection Agency (U.S. EPA, n.d.). NHD Plus data were initially developed at the 1:100,000 scale for the whole country. The current NHD Plus data is now developed at the 1:24,000/1:12,000 scale, which adds detail to the original dataset created at the 1:100,000 scale (U.S. EPA, n.d.). NHD Plus data were the hydrography data source utilized by SRI staff in their predictive modeling project.

It was also necessary to download four 10-meter resolution USGS DEMs covering the entire Hopi Reservation. These DEMs were downloaded from the USGS National Map Viewer and Download Platform. This data came from the National Elevation Dataset (NED) in a floating point pixel type and a GRID file format. These data were initially unprojected and had to be reprojected into the correct projection for analysis. The NED serves as the elevation layer of the National Map, and provides basic elevation information for earth science studies and mapping applications in the United States. The data extracted from this website are the elevation data utilized by the scientific and resource management communities for “research, hydrologic modeling, resource monitoring, mapping and visualization applications” (USGS, 2014). For this reason, this elevation dataset was utilized for creating topographic and water resource variables.

The two vegetation variables, vegetation richness with a 200- and 500-meter radii, were created from the GAP Land Cover Data Portal of the USGS. GAP data are a combination of several projects to create a seamless vegetation data set of the United States. The GAP data utilized for this project came from the Southwest Gap Analysis Project. These data were created from multi-season Landsat ETM+ from 1999 to 2001 in combination with DEMs “to model natural and semi-natural vegetation” (USGS, 2011). The USGS states:

> The GAP national land cover data, based on the NatureServe Ecological Systems Classification, are the foundation of the most detailed, consistent map of vegetative associations ever available for the United States and will help facilitate the planning and management of biological diversity on a regional and national scale (USGS, 2011).

This dataset was used in building models because it is the national standard when utilizing vegetation data in scientific studies and was also utilized by SRI in a similar predictive modeling project. These data originally came in a North American Datum (NAD) 83 Albers projection and had to be reprojected.

4.4 Data Collection Methods

There were no outside data collection efforts required for this project. All data were either provided by the client or downloaded from the web.

4.5 Data Scrubbing and Loading

There was extensive discussion and planning with the client on how the archaeological site type data were to be organized and scrubbed. The original archaeological point files provided by the client came in Microsoft Excel and Esri shapefile formats and in multiple projections. Before these datasets were loaded into the file geodatabase, the data in the attribute table had to be standardized. Initially, there were multiple archaeological site types. For simplicity and modeling purposes, the site types were combined into three types: Habitation, Rock Art, and Scatter. There were also multiple sites that were duplicated. This took time and coordination with the client to determine how to scrub the Esri shapefiles so that none of the archaeological sites were duplicated. There were also many naming convention errors for site types, which created problems when querying the data. For example, a Scatter site type was often called Lithic Scatter, and a Kiva site type had to be categorized under Habitation. Once all site types had been correctly categorized and combined into one Esri shapefile, the archaeological site types were all projected to the NAD 1927 UTM North Zone 12 projected coordinate system and then loaded into the file geodatabase in order to begin building models.

4.6 Summary

Chapter 4, Database Design, covered multiple topics of the data utilized for the predictive model for archaeological sites. The conceptual model addressed the underlying theory of how the predictive model for archaeological sites was developed. It addressed the three driving variables that determined archaeological site location: topography, water resources, and vegetation. This was addressed to clarify the logic behind the predictive model [Figure 4-1]. The conceptual model was then addressed to explain how the file geodatabase stored the necessary data and also the environmental raster variables were stored outside the file geodatabase. This was to discuss the logic of how the data for the model was stored and organized [Figure 4-2]. The data sources were then discussed not only to address where the data came from but also to justify why these datasets were used and how these were the best datasets for the predictive model. Scrubbing and loading of the data also were addressed to discuss the importance of how the data were initially acquired, organized, scrubbed and loaded into the file geodatabase.

Chapter 5 – Implementation

Chapter 5 discusses in detail how the project was implemented. The chapter was divided into eight sections to individually address each phase of project implementation. Section 5.1 discusses how the data were organized and prepared. Section 5.2 discusses independent variable development. Section 5.3 discusses how the data created in ArcGIS 10.2 was imported into SPSS. Section 5.4 discusses how the independent variables were tested for multicollinearity. Section 5.5 discusses how the models were built in SPSS. Section 5.6 discusses the model validation tests. Section 5.6 discusses how the predictive raster surfaces were created in ArcGIS 10.2. Section 5.7 discusses how the models were assessed for accuracy. Section 5.8 addresses how the predictive model surfaces were tested for accuracy. Section 5.9 summarizes the chapter and addresses the salient points regarding the overall implementation of the project. This was done in order for the client or another interested individual who is competent in GIS to be able to replicate the project.

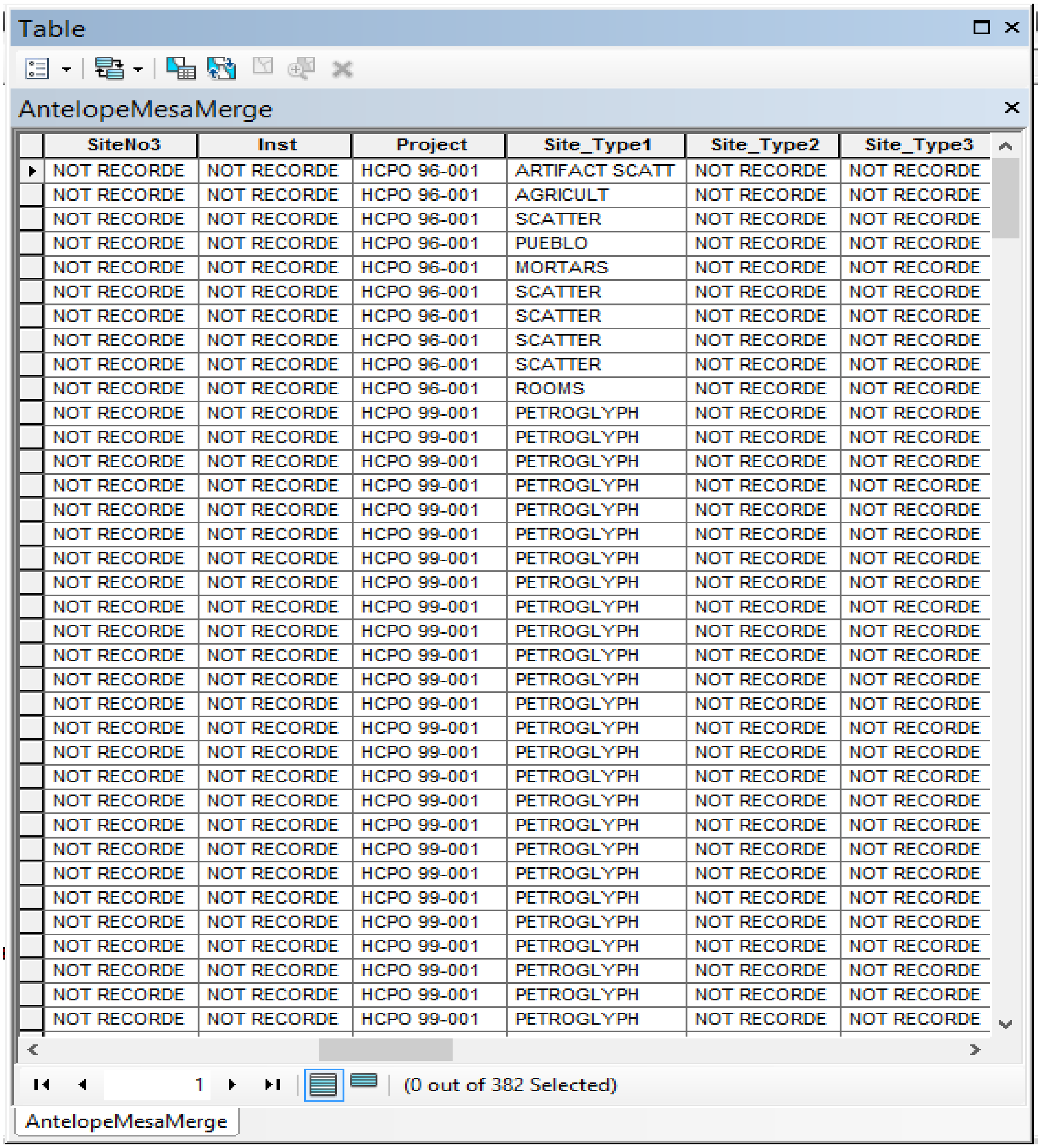
5.1 Data Compilation and Creation Methods

Archaeological site type data in Microsoft Excel and Esri shapefiles and the reservation polygon in Esri shapefiles were delivered directly by the client at the initiation of the project. This phase of the project required extensive exploratory analysis, organization, and scrubbing for the data to be ready to be input into the model. The site types included misspelled and redundant names and were not standardized in how they were entered into the attribute table [Figure 5-1]. The client requested that the following fields be included in the attribute table for the archaeological sites:

FID
SiteName
SiteNo1
SiteNo2
SiteNo3
Inst.
Project
SiteType1
SiteType2
SiteType3
Area
Paleo
Archaic
BM2
BM3
PI
PII
PIII
PIV

Historic
Depth
Residential

Some of the data were already populated in these fields, but a large amount of data had to be entered manually. For example, whether a site occurred during a particular archaeological period had to be identified by binary code (i.e., 1 = Yes and 0 = No). This was not done in the original attribute tables and had to be corrected. There was extensive discussion with the client in order for there to be a common understanding as to how the data were to be organized and stored. The client decided he was interested in modeling three archaeological site types: Habitation, Scatter, and Rock Art. Once all archaeological site type Esri shapefiles had been standardized and scrubbed, they were loaded into a single feature class called Archaeological Site Types.

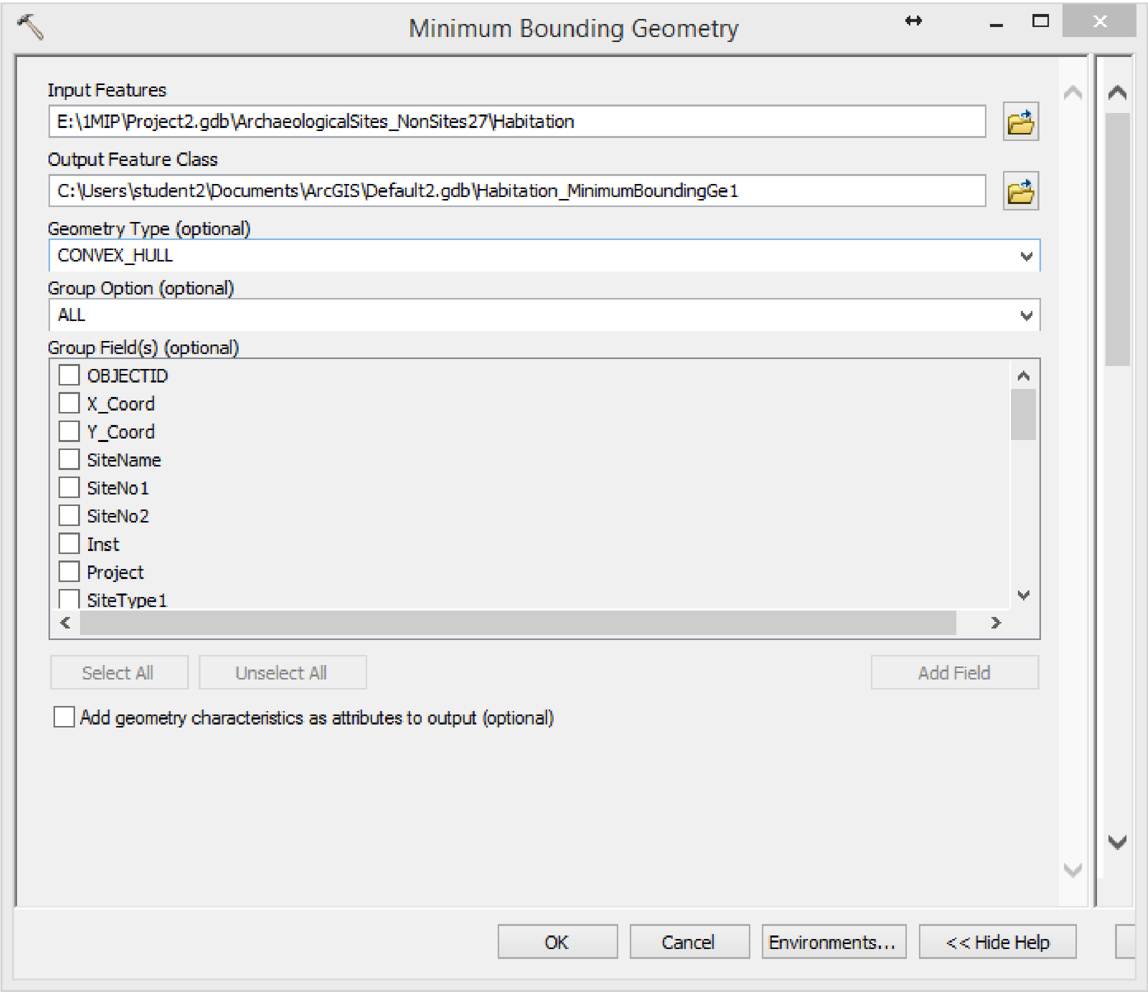


SiteNo3	Inst	Project	Site_Type1	Site_Type2	Site_Type3
NOT RECORDE	NOT RECORDE	HCPO 96-001	ARTIFACT SCATT	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	AGRICULT	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	SCATTER	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	PUEBLO	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	MORTARS	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	SCATTER	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	SCATTER	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	SCATTER	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	SCATTER	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 96-001	ROOMS	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE
NOT RECORDE	NOT RECORDE	HCPO 99-001	PETROGLYPH	NOT RECORDE	NOT RECORDE

Figure 5-1: The attribute table of archaeological sites displaying unstandardized site names.

Once the archaeological sites had been organized properly in accordance with the client's requests, the archaeological non-site types were created. Initially, the client wanted each archaeological site type modeled by the Basket Maker 3, Period 1, Period II, Period III, and Period IV archaeological periods. The issue arose of spatial dependency among sites. There was extreme clustering among site types in many parts of the Hopi Reservation. When spatial dependency occurs among features on the landscape, all the site types are occurring in very similar environmental conditions (e.g., slope, aspect, and cost to traverse to water). According to Goodchild, spatial dependency is "the propensity for nearby locations to influence each other and to possess similar attributes" (Goodchild, 1992). Consequently, the clustering of sites that are associated with the similar environmental conditions can lead to very strong biases in the model. Because of this, all the same site types (i.e. Rock Art, Habitation, and Scatter) were given 200-meter buffers, and boundaries were dissolved among the newly created buffers that overlapped. The 200-meter buffer distance was determined because with a larger buffer distance, many more sites would be eliminated which would not allow there to be a sufficient amount of sites to run the logistic regression analysis. According to Hosmer and Lemeshow, the general rule for the logistic regression model is that there should be 10 sites for every independent variable (Hosmer & Lemeshow, 2013). Because of this general rule by Hosmer and Lemeshow, the 200-meter buffer distance was chosen because if any larger there were not enough archaeological site points per independent variable to run the logistic regression model. Once the site types were re-created into dissolved buffers, the Feature to Point geoprocessing tool was run on all polygons created by buffers to generate polygon centroids. These were then the new feature classes for each archaeological site type to reduce spatial dependency among sites, which was previously causing the model to indicate bias and generate odd results.

Random sampling was then conducted to create archaeological non-site points to be combined with the site points in the logistic regression model. These point types were created to be the dependent, dichotomous variables in the model. The Minimum Bounding Geometry tool was run to create a convex hull polygon around the outermost points for each original archaeological site type dataset [Figure 5-2].



Minimum Bounding Geometry

Input Features

E:\1MIP\Project2.gdb\ArchaeologicalSites_NonSites27\Habitation

Output Feature Class

C:\Users\student2\Documents\ArcGIS\Default2.gdb\Habitation_MinimumBoundingGe1

Geometry Type (optional)

CONVEX_HULL

Group Option (optional)

ALL

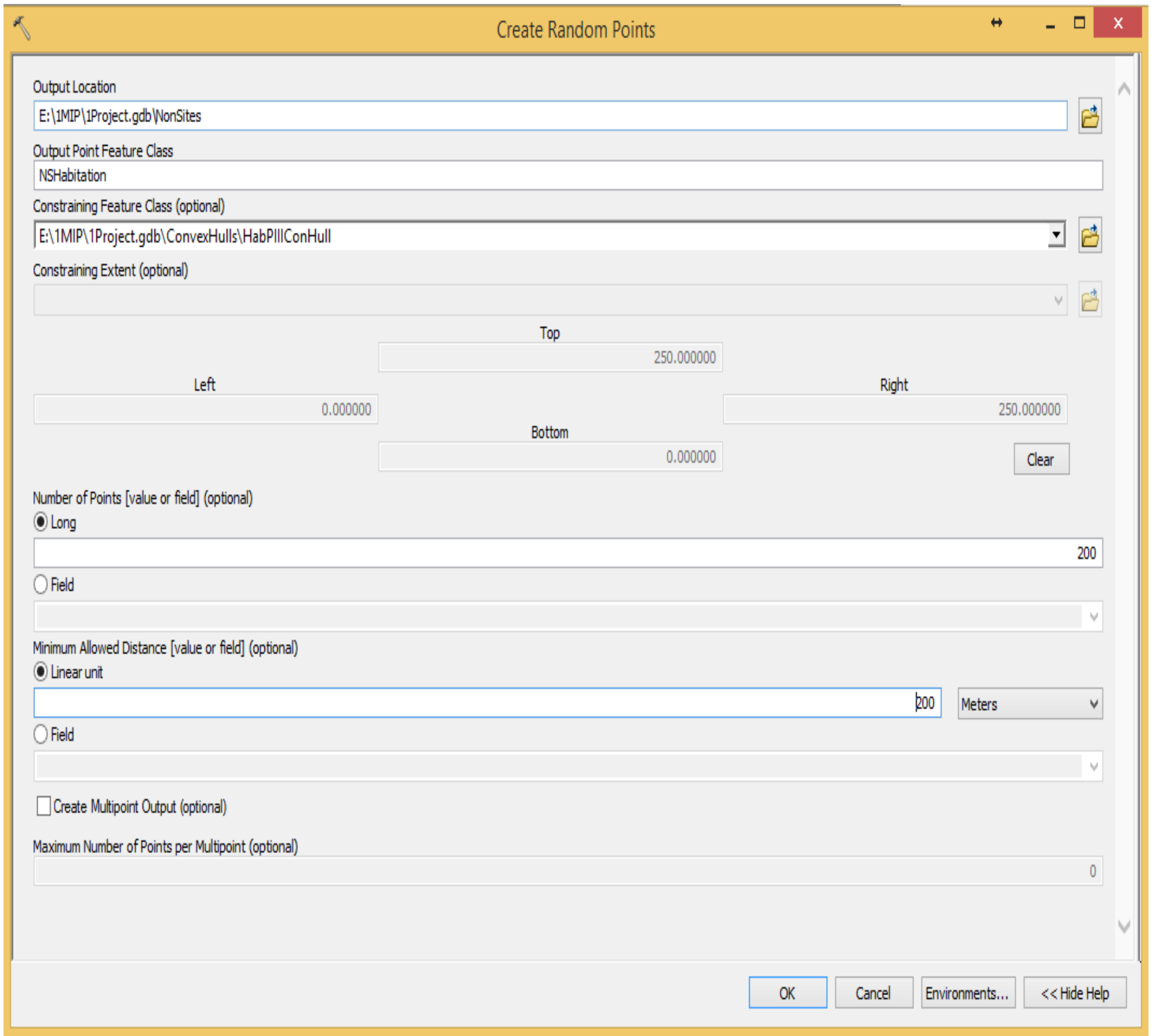
Group Field(s) (optional)

□ OBJECTID
□ X_Coord
□ Y_Coord
□ SiteName
□ SiteNo1
□ SiteNo2
□ Inst
□ Project
□ SiteType1

□ Add geometry characteristics as attributes to output (optional)

Figure 5-2: The Minimum Bounding Geometry geoprocessing tool to create convex Hulls.

This was done to create the smallest sampling polygon possible that contained all the archaeological site types. The Create Random Points geoprocessing tool was then run to create close to the same amount of non-sites as sites for each category [Figure 5-3]. All points that were outside the reservation and within 200-meters of each other were eliminated. This was imperative to try to eliminate the possibility of creating a random non-site point that intersected another site point or else more errors would be introduced into the model.



Output Location
E:\1MIP\1Project.gdb\NonSites

Output Point Feature Class
NSHabitation

Constraining Feature Class (optional)
E:\1MIP\1Project.gdb\ConvexHulls\HabPIIIConHull

Constraining Extent (optional)

Top
250.000000

Left
0.000000

Right
250.000000

Bottom
0.000000

Number of Points [value or field] (optional)
● Long
200
○ Field

Minimum Allowed Distance [value or field] (optional)
● Linear unit
200 Meters
○ Field

□ Create Multipoint Output (optional)

Maximum Number of Points per Multipoint (optional)
0

Figure 5-3: The Create Random Points geoprocessing tool to create non-site points within the convex hull.

Once an appropriate amount of non-site points were created for each archaeological site type, the independent variables were then created. The topographic and water resource variables were created from a 10-meter USGS DEM. Initially, multiple DEMs were mosaicked together using the Mosaic by Mask geoprocessing tool. When this was completed there was a noticeable gap that was created when the USGS DEMs were mosaicked together that contained null values [Figure 5-4].

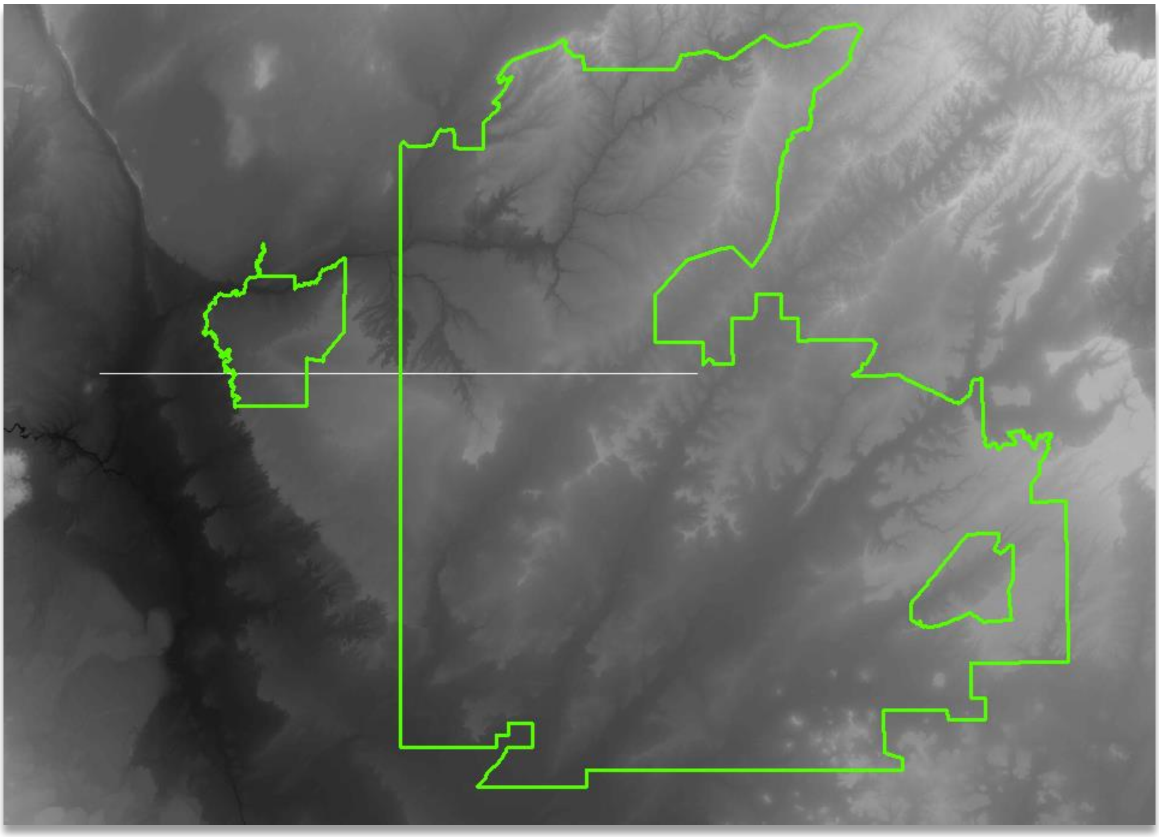


Figure 5-4: Large gap in the mosaicked USGS DEM

Because this null value gap could introduce many errors in the data when used in raster analysis, it was necessary to replace the null values in the USGS DEM by applying the following formula in the Raster Calculator geoprocessing tool to this mosaicked raster:

> Con(IsNull(“raster”), FocalStatistics(“raster”, NbrRectangle(27,27,), “MEAN”), “raster”)

where in the conditional statement, if the raster pixel is a null value the Focal Statistics tool determined the mean elevation values of the neighboring pixels that are not null. In order to completely fill in the gaps, 27 pixels in height were determined to be the widest gap in the newly mosaicked USGS DEM. This number was determined by trial and error until the gap was completely removed. By selecting 27, the Focal Statistics geoprocessing tool was able to completely fill in all the null values in the gap because 27 was the largest gap in height. The Focal Statistics geoprocessing tool was then able to apply the mean elevation of the neighboring pixels to the null values in the gap.

5.2 Independent Variable Development

Once this problem was successfully addressed, the USGS DEM was ready to be used for the creation of the topographic and water resource independent variables. The topographic independent variables consisted of the following:

- Slope
- North-South Aspect
- East-West Aspect
- Local Elevation Change
- Terrain Texture
- Shelter
- Cost Surface
- Topographic Position Index

These were all created in ArcGIS 10.2 using various geoprocessing tools. Slope is known to be one of the most common explanatory variables used in predicting site locations (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Percent rise in slope was calculated with the Slope tool. Neighboring pixels tended to have very different slope values. This was typically due to pixels being either on, above, or below contour lines used to create the DEM (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). In order to smooth this extreme variation in slope values, the Focal Statistics tool was utilized to calculate the mean percent slope within a 100-meter radius [Figure 5-5].

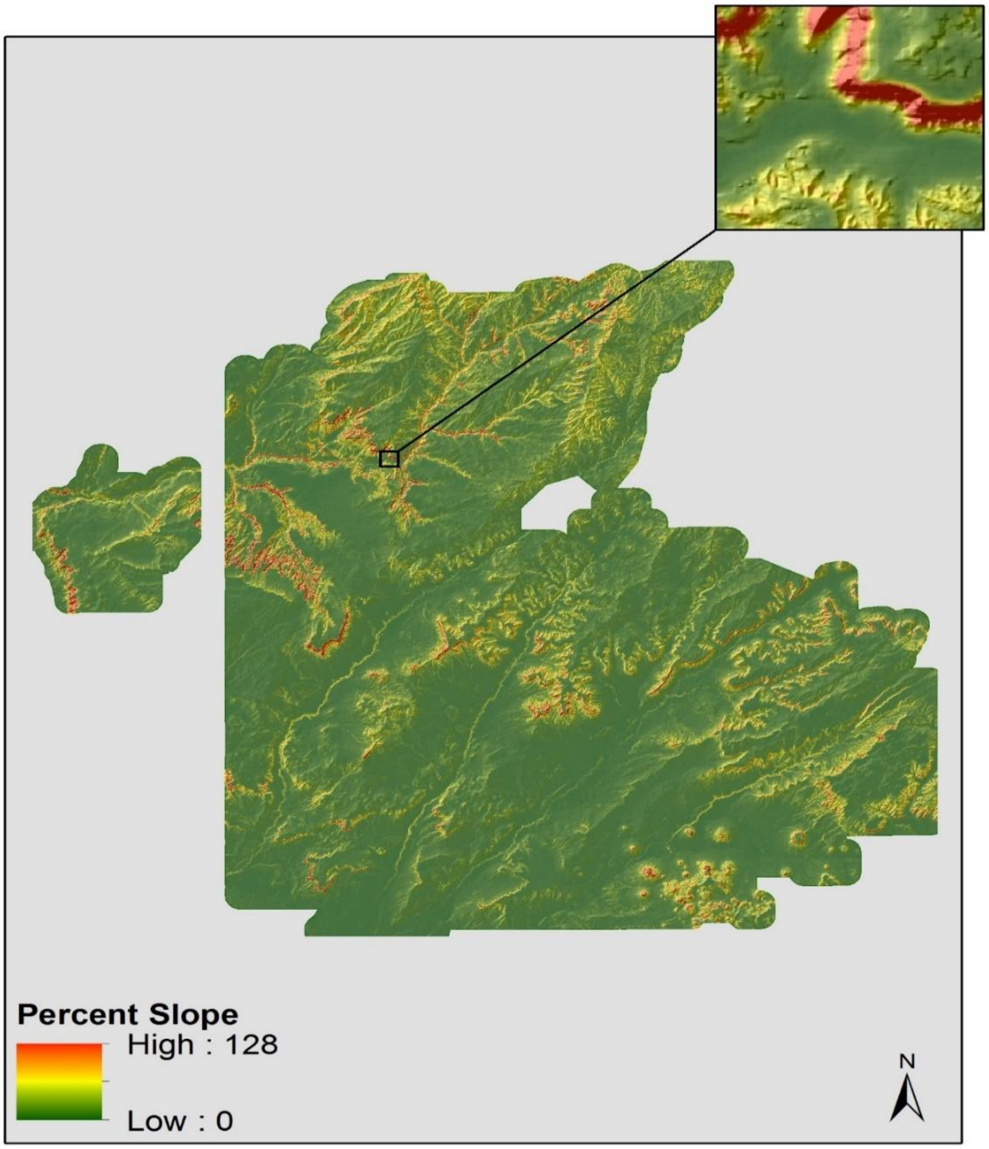


Figure 5-5: The slope variable created, which can influence site location.

Aspect was another variable that was used in model development [Figures 5-6 and 5-7]. Aspect is important, for example, because a southern slope can have excellent sun exposure or provide great protection from strong winds (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Aspect was created using the Aspect geoprocessing tool and the USGS DEM as the input raster. The inherent problem of aspect when calculated in ArcGIS is that it corresponds to the degrees of a compass. The problem herein lies in that 359^o and 1^o may be quantitatively distinct, although, there is very little difference in these directions in reality (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). In order to address this problem, Heilen et al. (2012) rescaled aspect so it ranged from 0^o to 180^o. By doing this, aspect "was then distributed symmetrically along either a north-south or an east-west axis" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). This allowed aspects to be generalized in either a north-south or east-west fashion. The following equation in the Raster Calculator was applied to the newly created aspect raster to transform all northerly directions to equal 0º and all southerly directions to equal 180º:

```
Con(("aspect" < 0),90,Con(("aspect" > 180),(360 - "aspect"),"aspect"))
```

where in the first conditional statement, any aspect degree less than 0º was given 90º. This indicated that any pixel that was 90º was not trending either north or south. In the second conditional statement, when there was an aspect degree greater than 180º, the aspect raster was subtracted from 360º. This standardized these aspect values greater than 180º, and converted them to values less than 180º. For example, 181º was then converted to 179º and a pixel that was already 179º, maintains that value because it is already less than 180º. The final statement in the algorithm, “aspect,” maintained that all aspect values that were less than 180º stay the same that were originally created in the original slope raster. The following formula was also applied in the Raster Calculator to generalize all east-west directions:

```
Con(("aspect" < 0),90,Con(("aspect" < 90),(90-"aspect"),Con("aspect" <
270,("aspect" - 90),(270 - ("aspect" - 180)))))
```

where in the first conditional statement, any aspect degree less than 0º was given 90º, which is the same as the previous formula. In the second conditional statement, any value less than 90º was subtracted from 90º. For example, 15º is now 75º when it is deducted from 90 º. In the third conditional statement, 90º was subtracted from all values less than 270º and greater than or equal to 90º. For example, if 90º was subtracted from 260º, the new values is now 170º. This is now less than 180º. The last part of the third conditional statement subtracts 180º from all values greater than 270º. This formula transformed all easterly directions to be between 0º and 89.9º all westerly directions to be between 90.1º and 180º (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012).

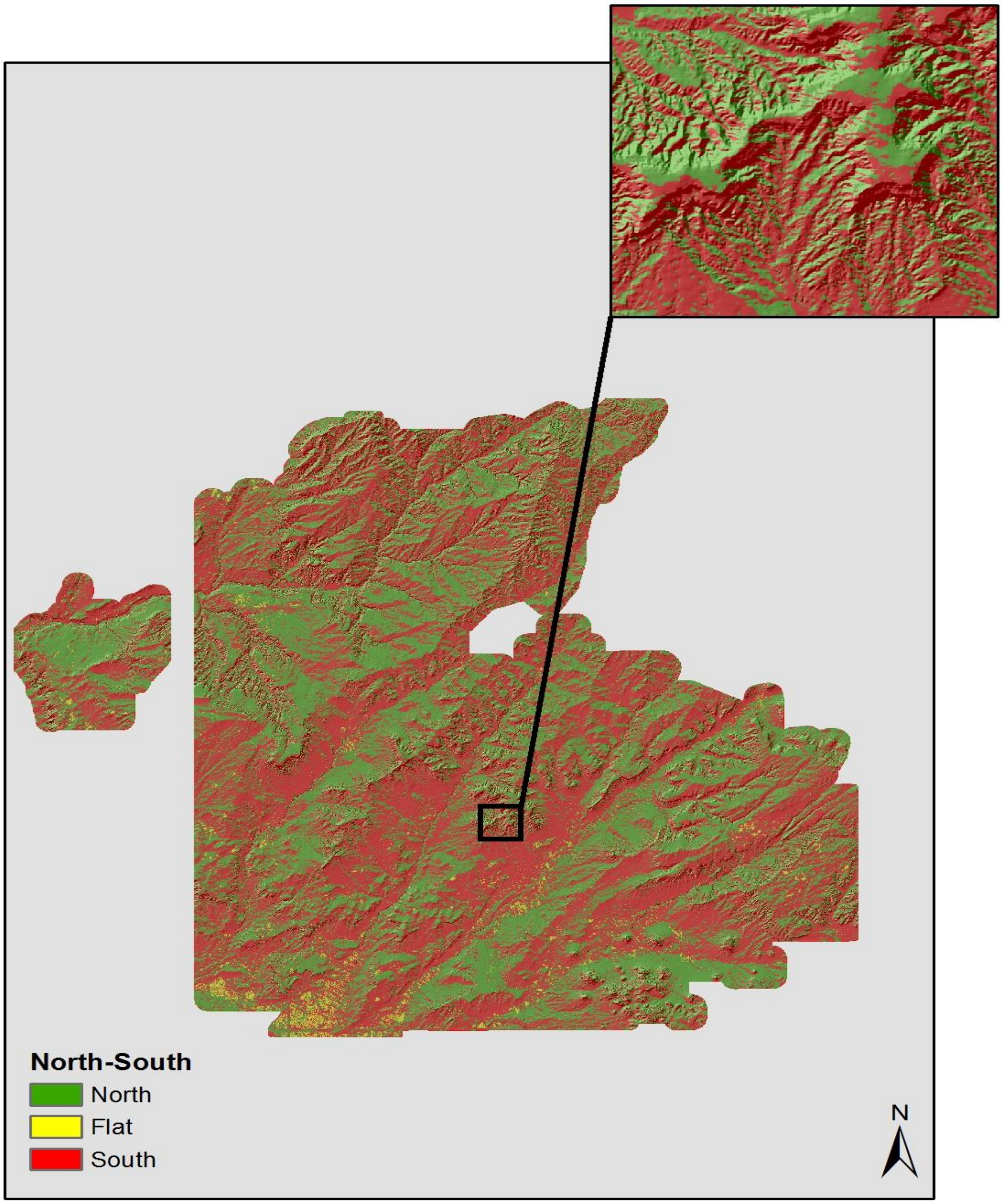


Figure 5-6: The north-south aspect variable which can influence site location.

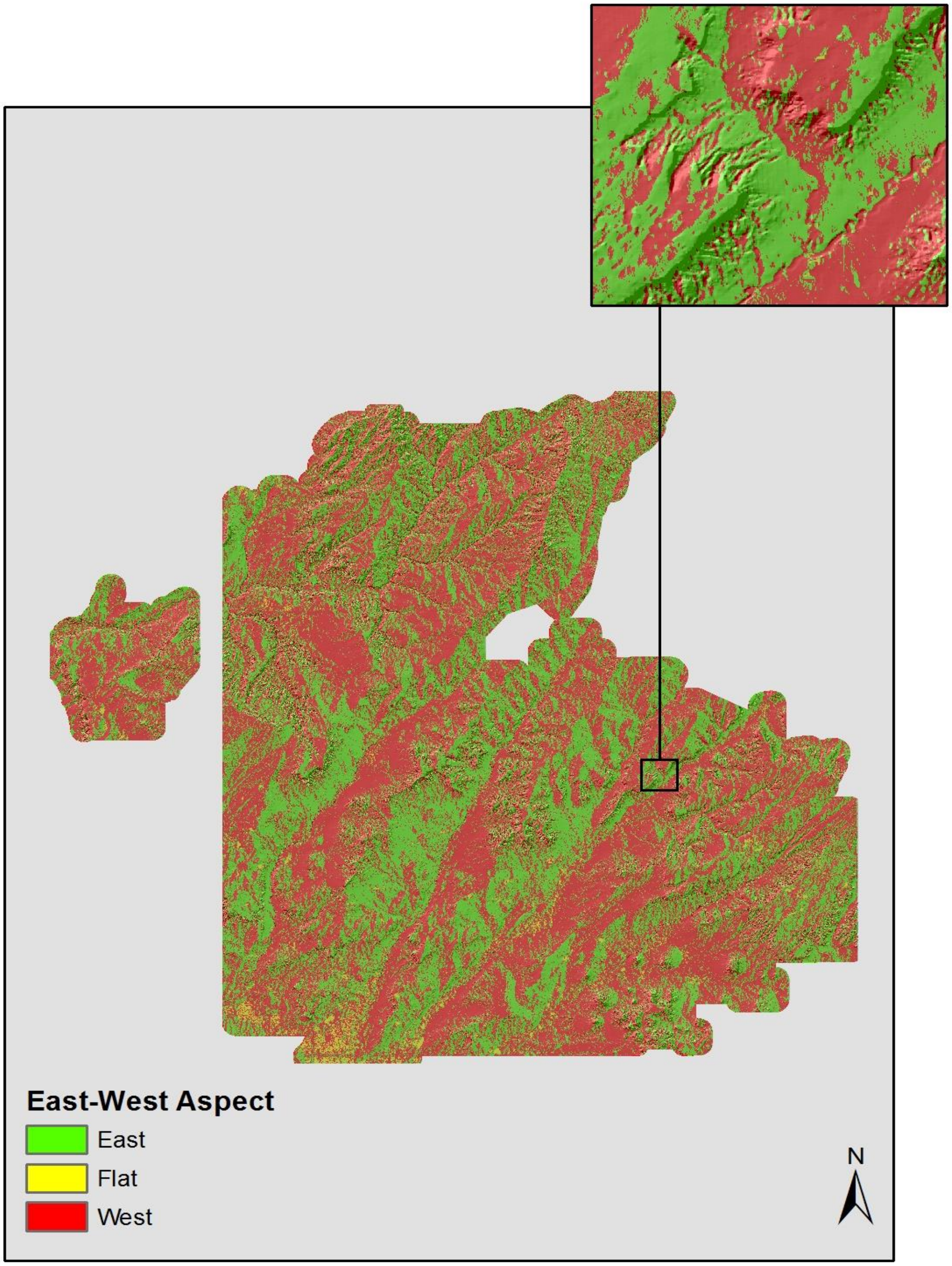
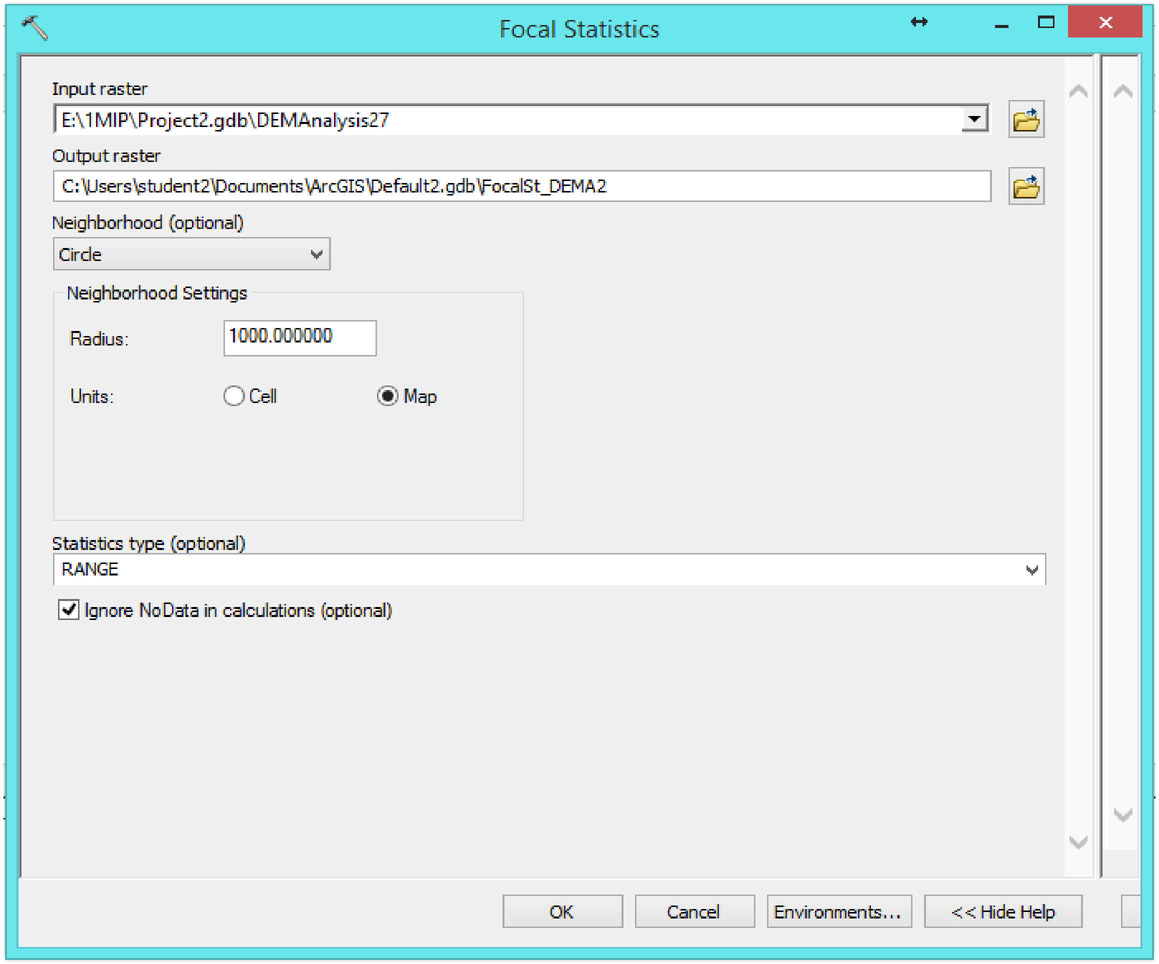


Figure 5-7: The east-west aspect variable which can influence site location.

Local elevation change was the next topographic variable utilized in building models. Local elevation change is a measure of roughness and is important in dictating site location (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Local elevation change is considered to be a variable that has an impact on the ease of travel across landscapes. According to SRI, local elevation change is “the range in elevation within a predefined radius around a raster cell” (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Local elevation change with a 1-kilometer radius was calculated with the Focal Statistics geoprocessing tool and was used in building models [Figures 5-8 and 5-9].



Input raster
E:\1MIP\Project2.gdb\DEMAnalysis27

Output raster
C:\Users\student2\Documents\ArcGIS\Default2.gdb\FocalSt_DEMA2

Neighborhood (optional)
Circle

Neighborhood Settings

Radius: 1000.000000

Units: ○ Cell ● Map

Statistics type (optional)
RANGE

☑ Ignore NoData in calculations (optional)

Figure 5-8: The Focal Statistics geoprocessing tool utilized to create the local elevation change variable.

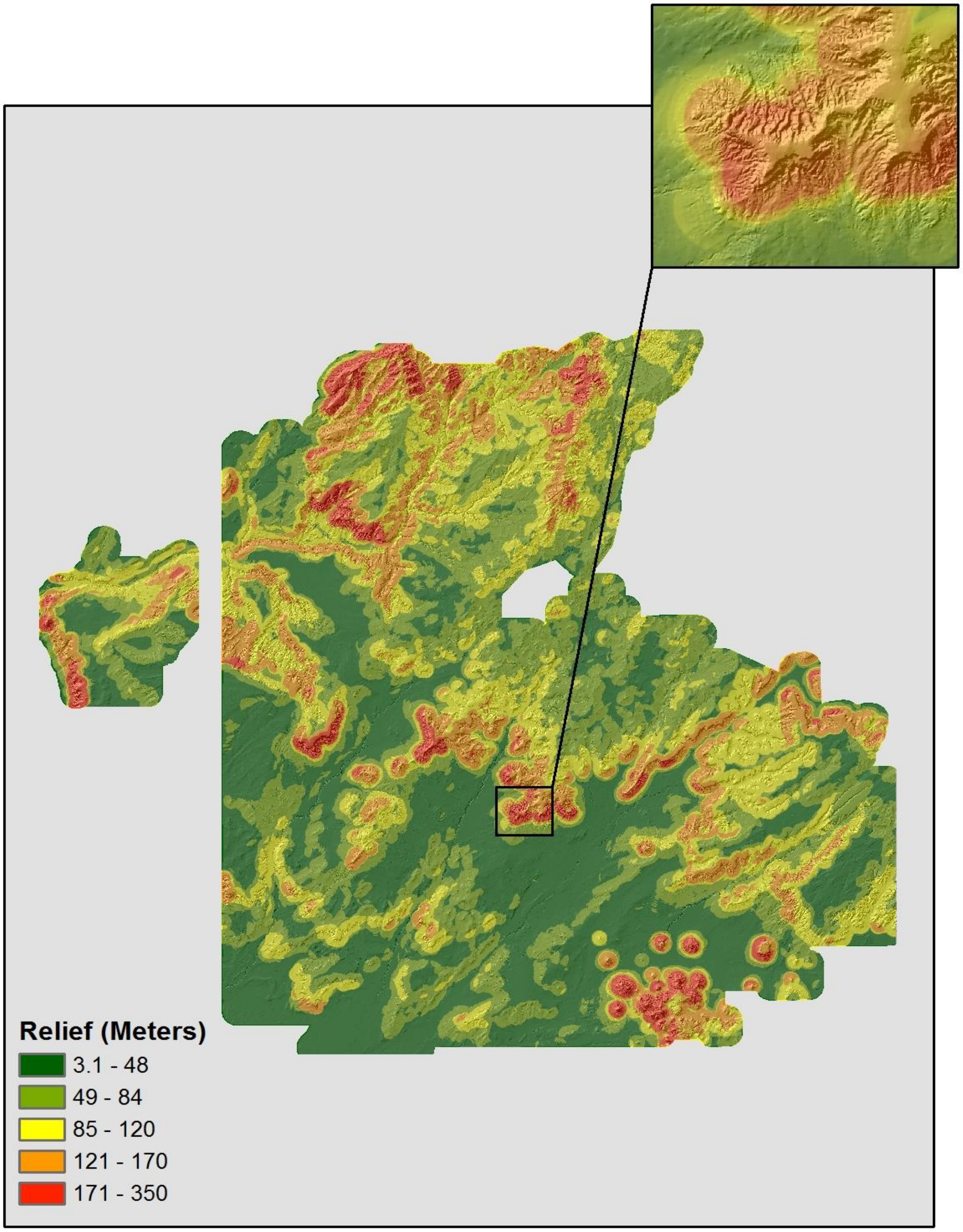


Figure 5-9: The local elevation change variable, which measures surface roughness, an indicator of how easy it is to traverse the landscape.

Terrain texture was another variable used in building models [Figure 5-10]. Terrain texture “is the amount of variability in elevation within a predefined radius” and is considered another way to measure surface roughness (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). This variable was created by calculating the standard deviation in elevation with a radius of 1 kilometer with the Focal Statistics geoprocessing tool. The higher the standard deviation indicated a higher level of surface roughness or terrain texture.

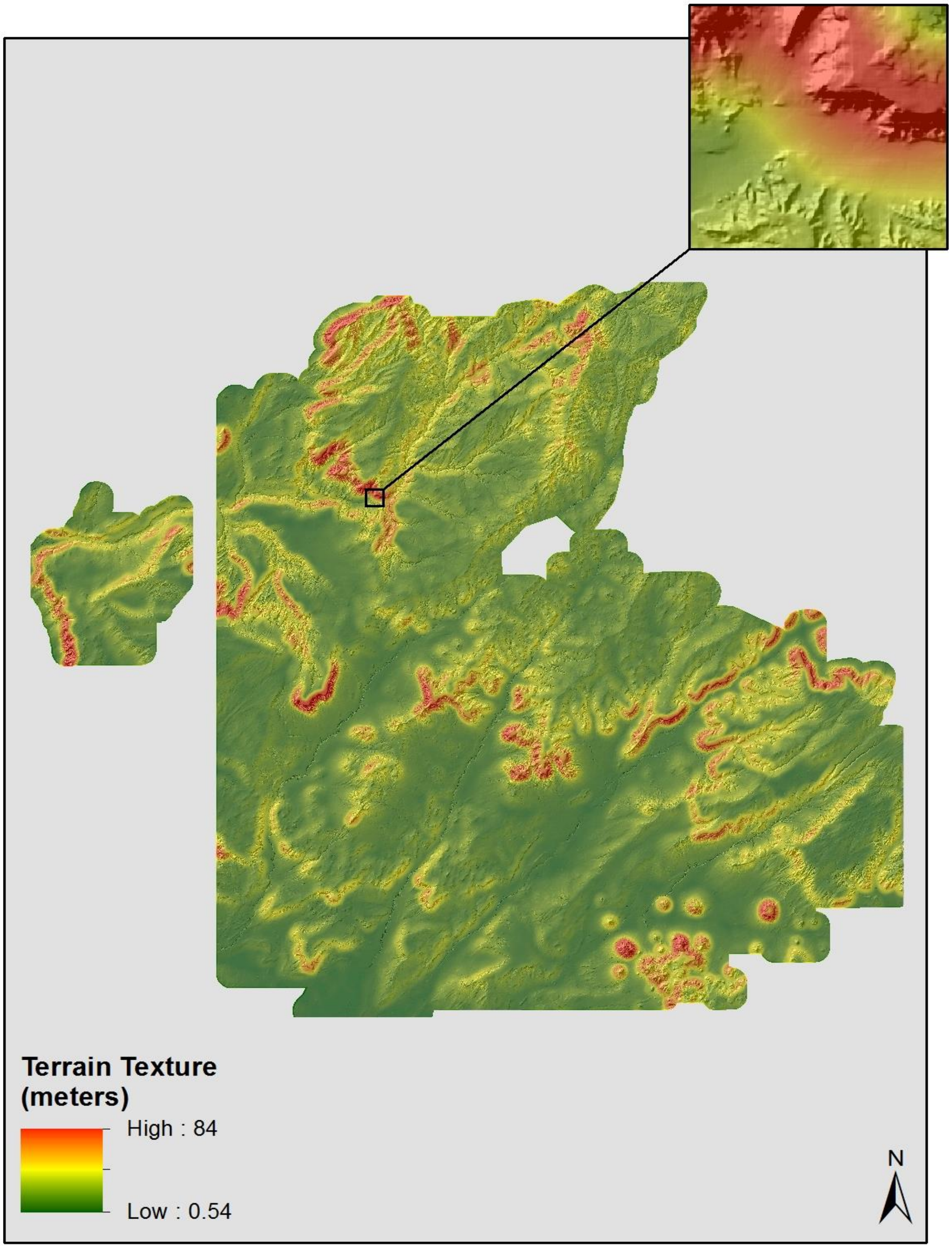


Figure 5-10: The terrain texture variable created to measure the amount of variability within a 1-kilometer radius.

Shelter was another variable that was found to be important in building models [Figure 5-11]. This variable explains the degree to which the surrounding topographic features offer shelter from the environmental elements (e.g., sun, wind) (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Low-lying areas surrounded by hills are considered to provide a high level of shelter. To create this variable the mean elevation within a 1-kilometer radius was created with the Focal Statistics geoprocessing tool. Shelter was then calculated in the Raster Calculator by dividing "the mean elevation within a specified radius by the local elevation of a given raster cell" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Features on the landscape with values above 1 indicate more shelter, while features corresponding to values less than 1 represent areas with less shelter.

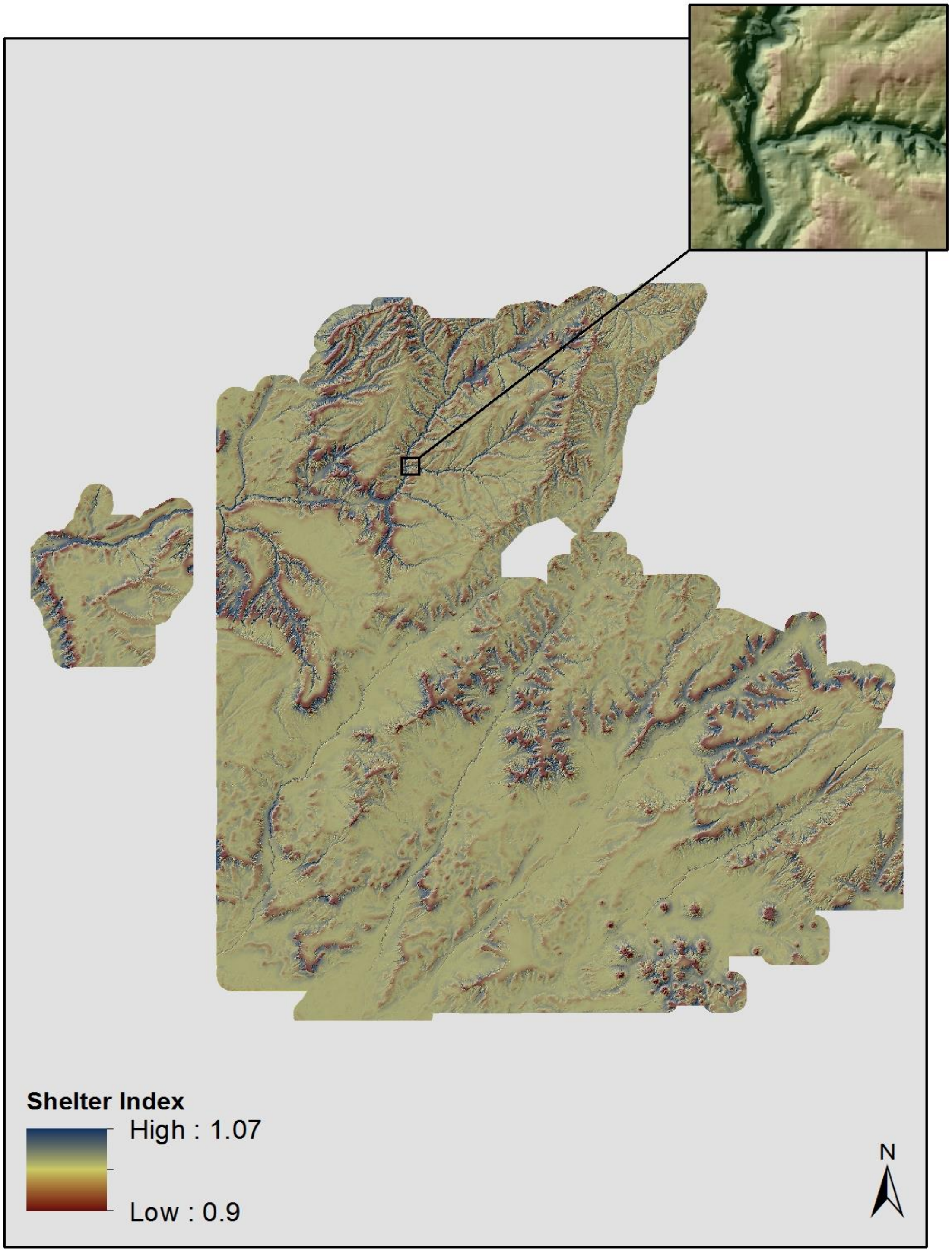


Figure 5-11: The shelter variable created to measure the amount of protection the landscape can provide sites.

Topographic Position Index (TPI) was another topographic variable utilized in building models [Figure 5-12]. This variable addressed where a site was located or positioned on the landscape in regard to high points such as ridges and low points such as valley floors. According to Weiss, TPI “compares the elevation of each cell in a DEM to the mean elevation of a specified neighborhood around that cell” (Weiss, 2014). Features on the landscape corresponding to values closer to 1 represent high locations, while features corresponding to values closer to 0 represent lower-lying areas. TPI was calculated by creating a focal minimum elevation raster and a focal maximum elevation raster with the USGS DEM and using the Focal Statistics geoprocessing tool. The following equation was then applied to these newly created rasters in the Raster Calculator to create the TPI variable (Cooley, 2014):

> (DEM – Focal minimum elevation)/(Focal maximum elevation – Focal minimum elevation)

As with the variables previously mentioned, TPI can help explain site location, whether the past inhabitants wanted to settle near valley floors, more exposed areas with better views, or areas with better protection with higher surface roughness.

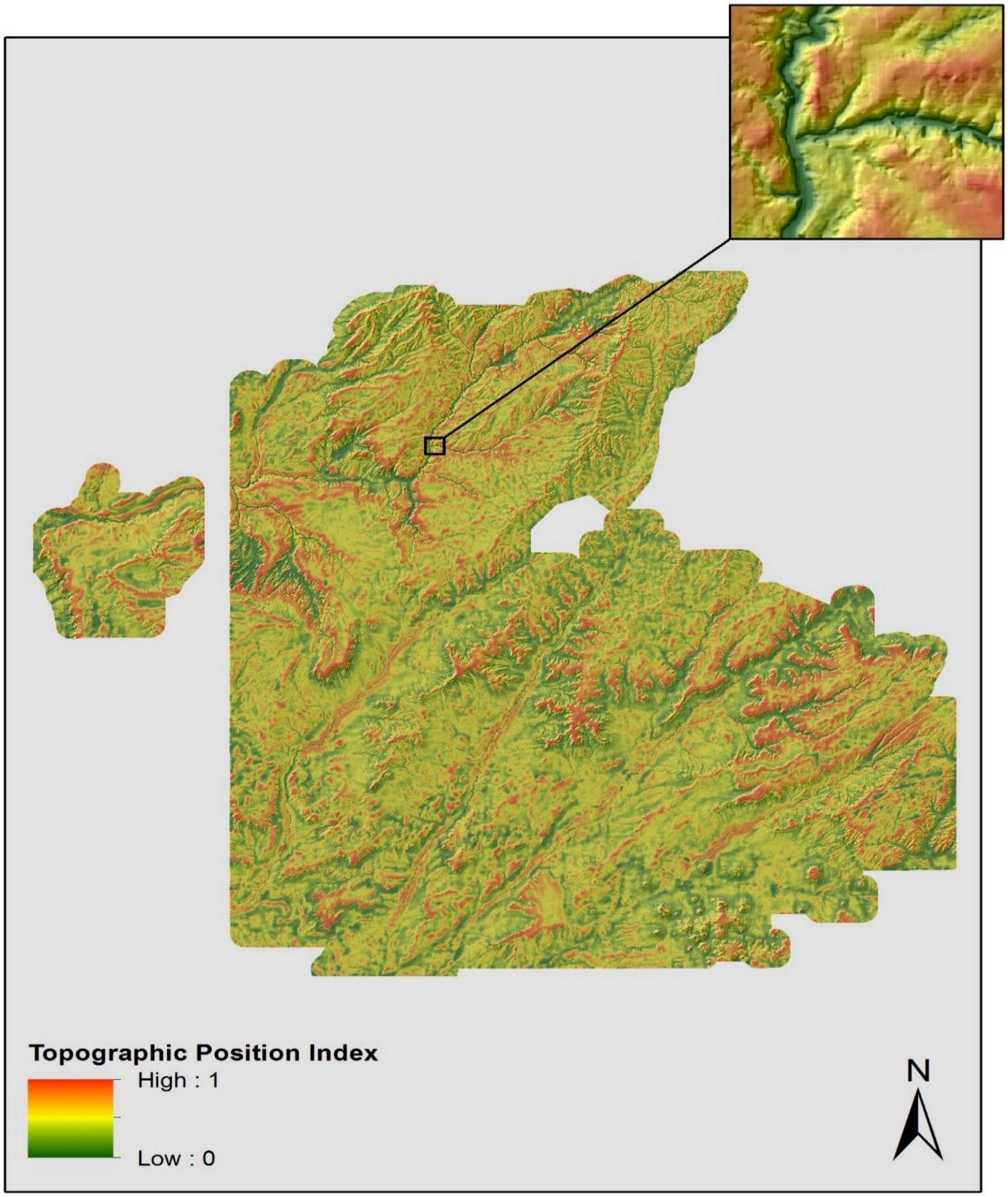


Figure 5-12: The TPI variable created to explain a sites location on the landscape.

Cost to traverse was the final topographic variable calculated [Figure 5-13]. This was one of the most important variables utilized in building models because it was the basis for how all the water resource variables were calculated. It can be used to "measure the [cost to traverse] between a given raster cell" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012) and a resource type (e.g., water). This variable "represents the relative cost of moving across the landscape" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). The following methodology was utilized to calculate a cost to traverse variable:

1. The mean slope within a 100-meter radius of the DEM and the standard deviation within a 100-meter radius of the same original DEM were calculated with the Focal Statistics geoprocessing tool.
2. A value of 1 was then added to each of these newly created variables with the Raster Calculator.
3. The natural logarithm of each variable was then calculated with the Ln geoprocessing tool
4. Add the 2 resultant values together with the Raster Calculator to create the final cost to traverse independent variable (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012)

This variable addressed both the slope and the ruggedness of a specified radius surrounding a raster cell (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Areas with the steepest slopes have the highest cost to traverse, whereas areas with the lowest slopes have lower cost to traverse (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012) [Figure 5-13].

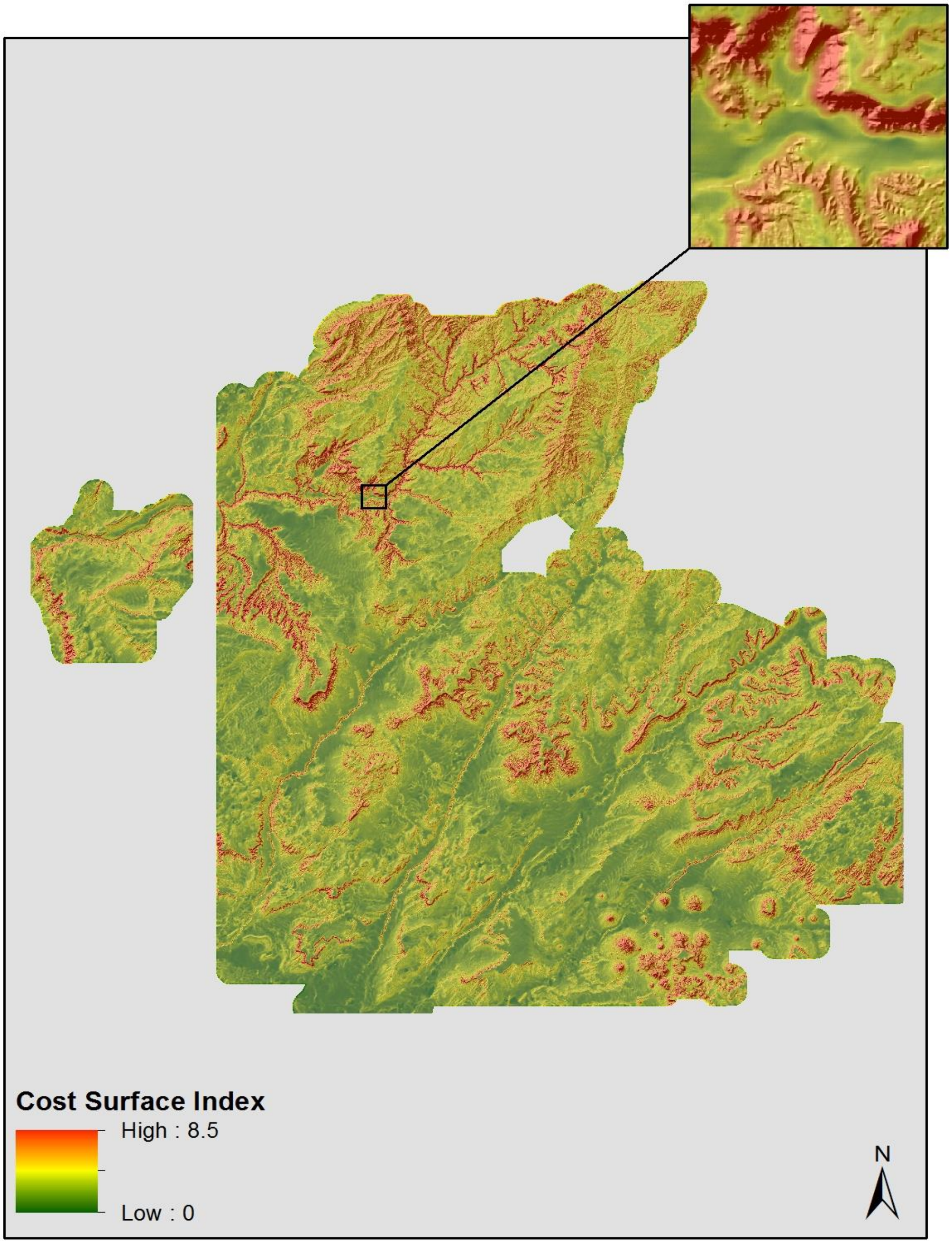


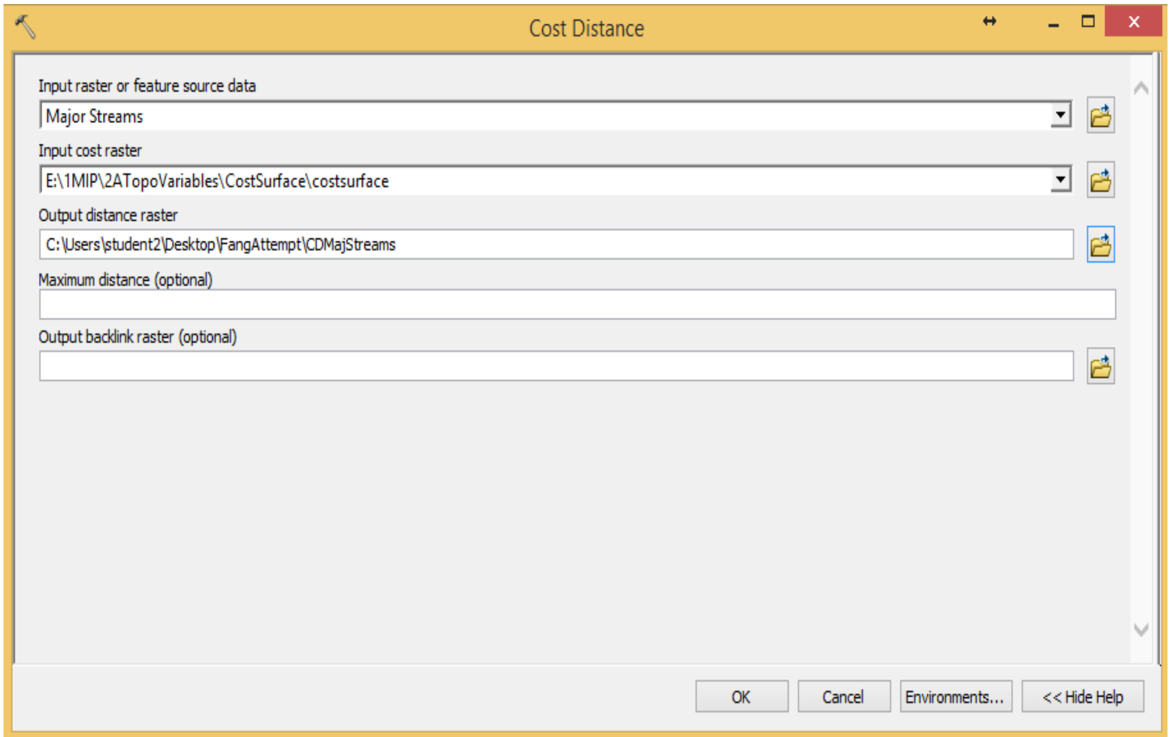
Figure 5-13: The cost to traverse variable created to measure the relative difficulty of moving across the landscape.

Water resources were the next variables considered when considering independent variables for predictive modeling. There were 6 variables created in total. These were:

- Cost to traverse to major to streamlines
- Cost to traverse to all streams
- Cost to traverse to springs or seeps
- Cost to traverse to water bodies
- Cost to traverse to minor stream network nodes
- Cost to traverse to major stream network nodes

These were created in ArcGIS 10.2 with various geoprocessing tools.

Cost to traverse to major streamlines was created by using NHD Plus stream lines data. Canals and pipelines were removed from this original feature class in order to analyze only naturally occurring water resources. The Strahler stream order classification methodology was implemented to determine the highest-order streams within the reservation (Strahler, 1957). Stream orders 3 through 5 were identified as major streams within the reservation. The cost to traverse to major stream lines variable was then created by using the Cost Distance geoprocessing tool with the major stream lines feature class and the cost to traverse variable described above [Figures 5-14 and 5-15].



Input raster or feature source data
Major Streams

Input cost raster
E:\1MIP\2ATopoVariables\CostSurface\costsurface

Output distance raster
C:\Users\student2\Desktop\FangAttempt\CDMajStreams

Maximum distance (optional)

Output backlink raster (optional)

Figure 5-14: The Cost Distance geoprocessing tool used to create the cost to traverse to major streams variable.

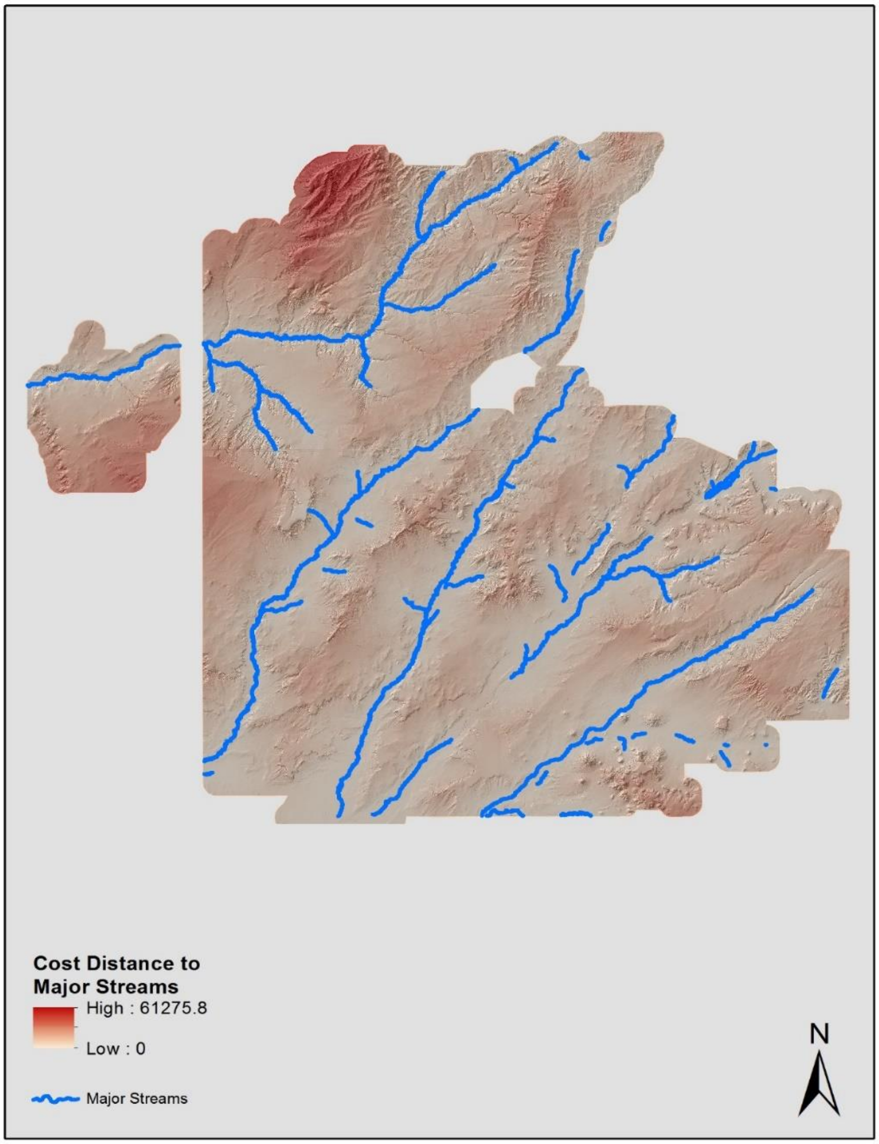


Figure 5-15: The cost to traverse to major streams variable created to measure the difficulty level to traverse the landscape to this water resource.

The cost to traverse to all streams variable was also created by using NHD Plus stream lines data. The cost to traverse to stream lines variable was calculated with the Cost Distance geoprocessing tool by using this stream lines feature class and the cost to traverse variable described above [Figure 5-16].

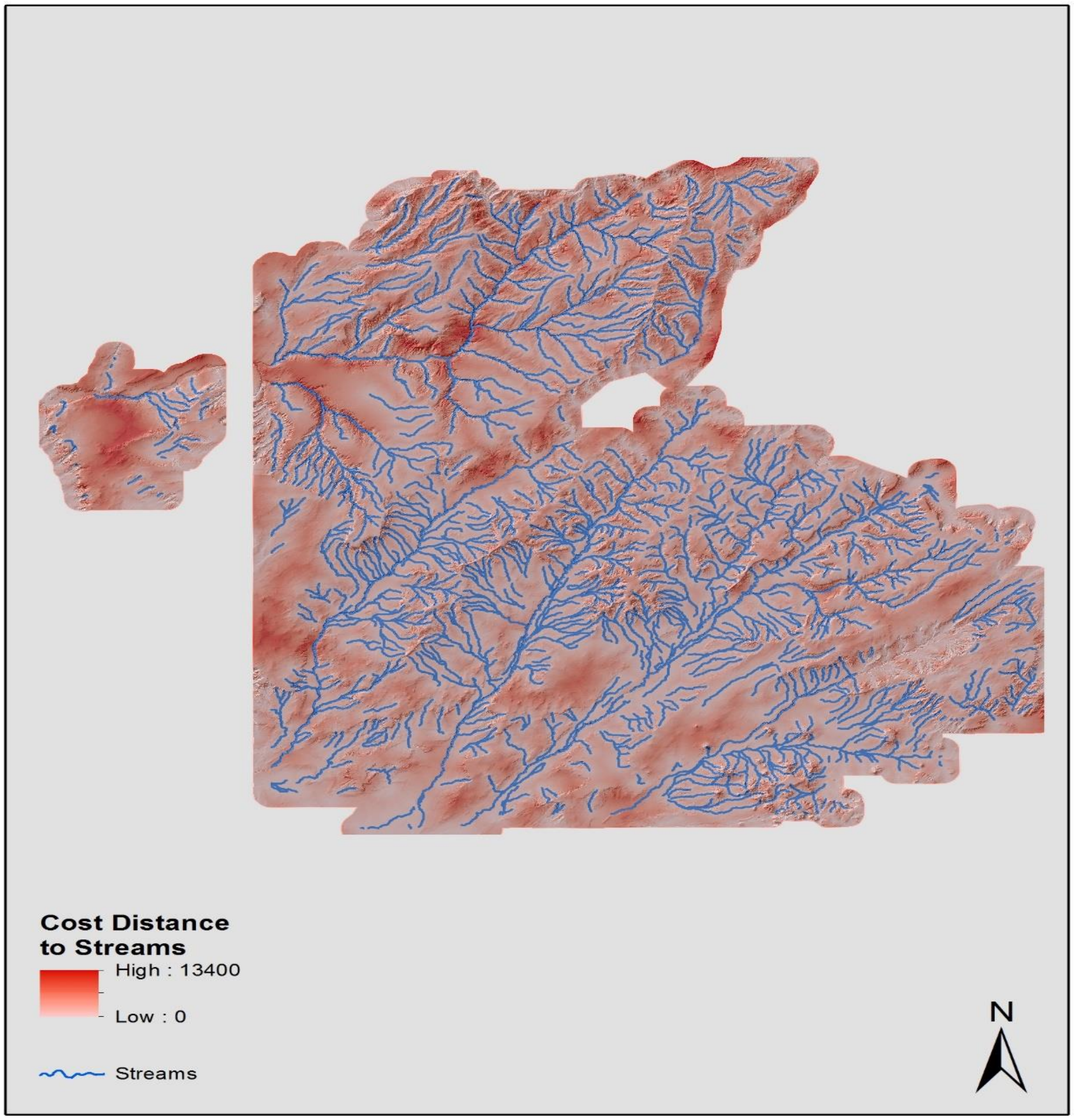


Figure 5-16: The cost to traverse to all streams variable created to measure the difficulty level to traverse the landscape to all streams, excluding the major streams.

NHD Plus data was also used for analyzing the cost to traverse to springs and seeps. This variable was created using the Cost Distance geoprocessing tool with the springs and seeps feature class and the cost to traverse variable [5-17]. The cost to traverse to water bodies and cost to traverse to stream network nodes were created by following the same procedure [Figures 5-18 and 5-19]. The major stream network nodes were considered to be "stream junctions that were situated along 'major' stream segments…." (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). The major stream network nodes were extracted by utilizing the Feature Vertices to Points geoprocessing tool and additional visual inspection to ensure that all necessary stream junction nodes were extracted. The same procedure as above was followed with the major stream network nodes as the input feature source data to create the cost to traverse to major stream network nodes [Figure 5-20].

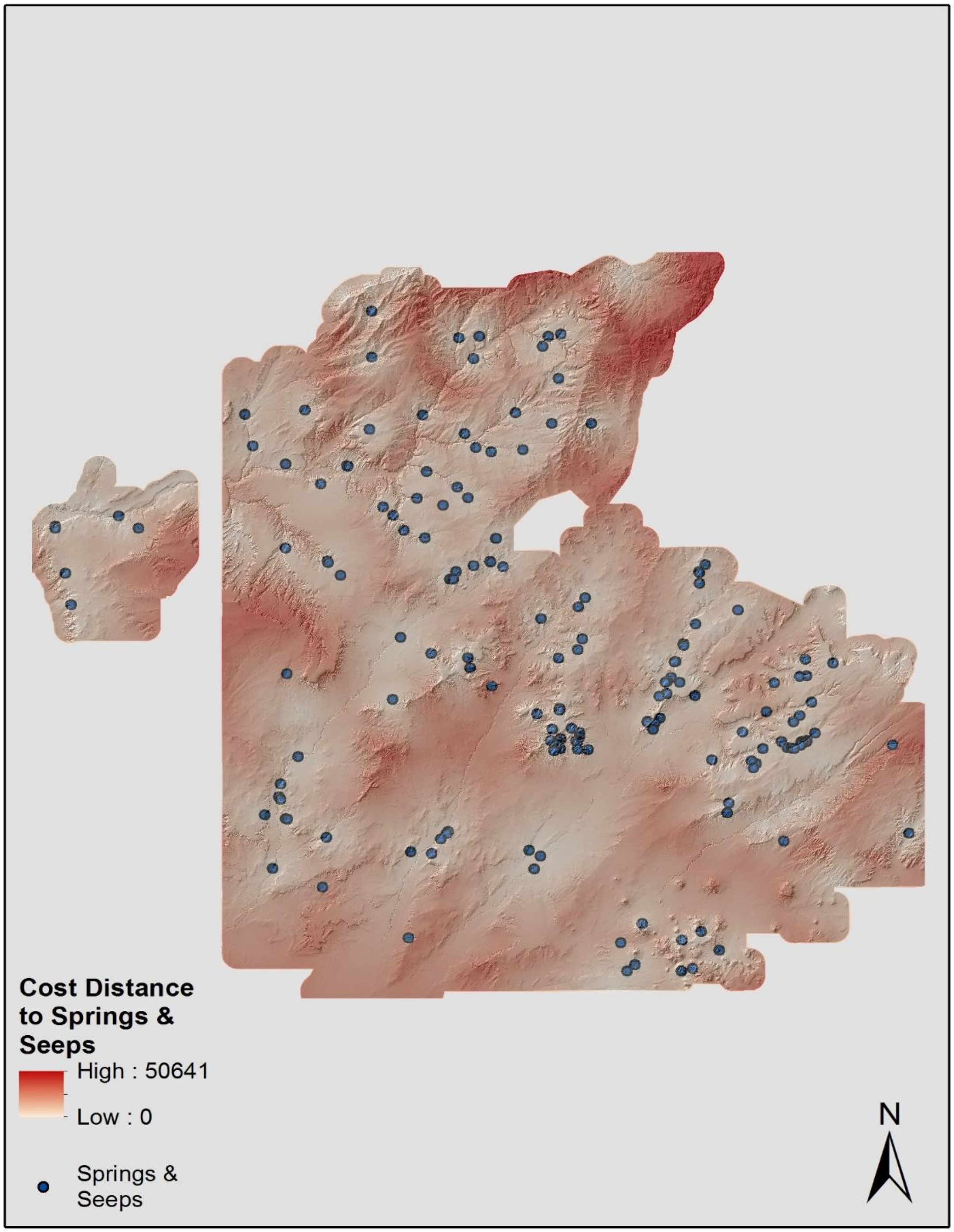


Figure 5-17: The cost to traverse to springs and seeps variable created to measure the difficulty level to traverse the landscape to this water resource.

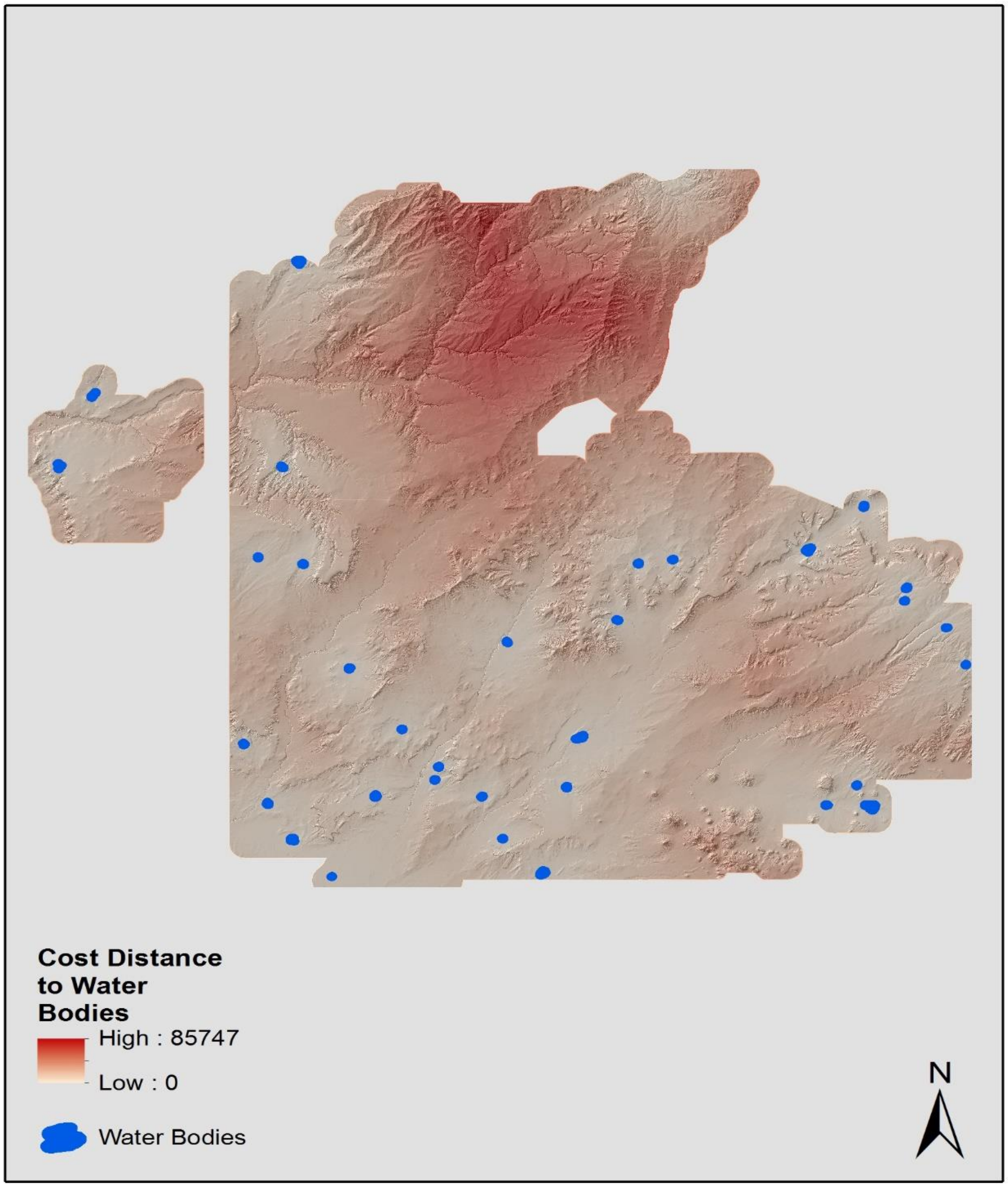


Figure 5-18: The cost to traverse to water bodies variable created to measure the difficulty level to traverse the landscape to this water resource.

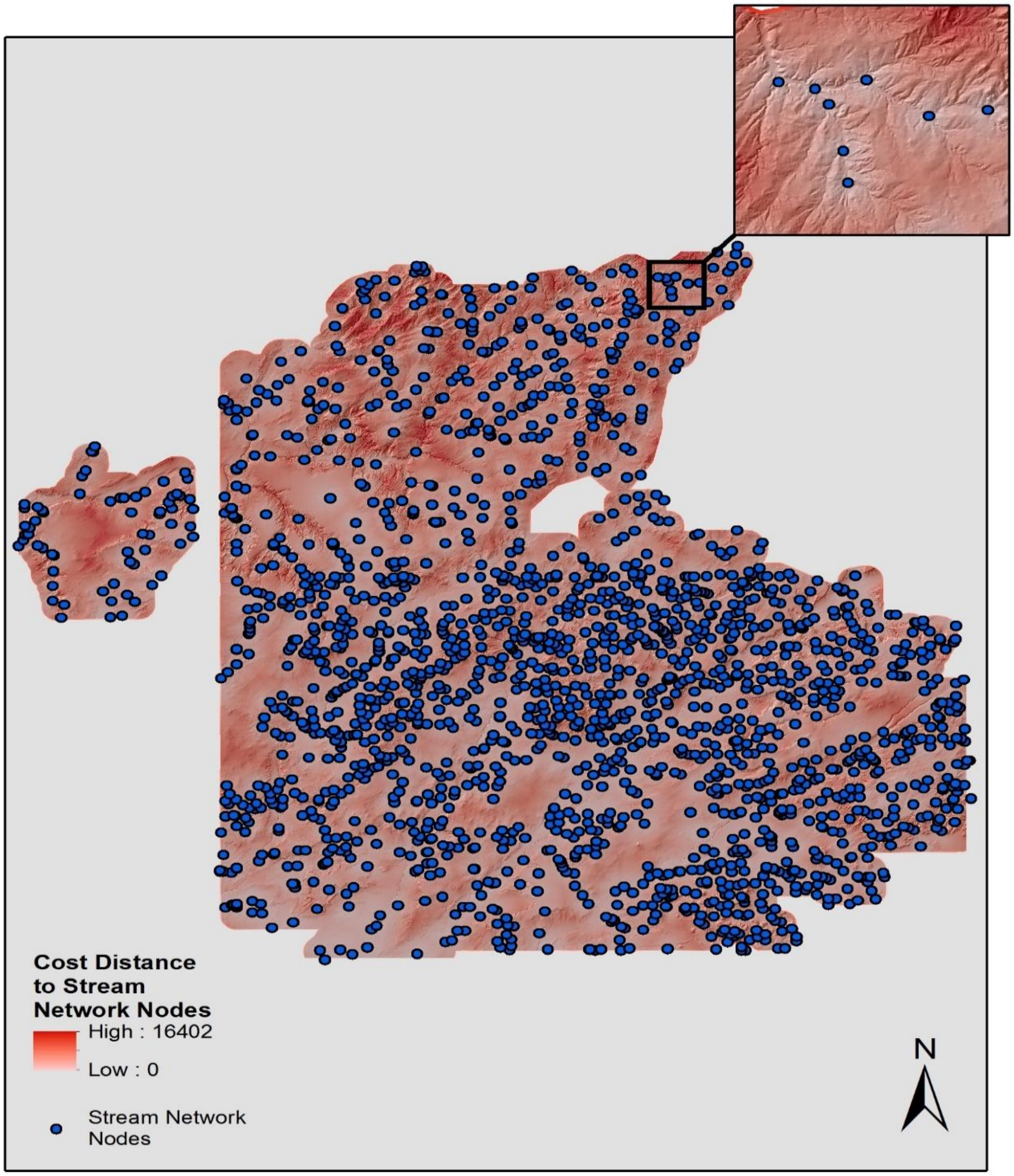


Figure 5-19: The cost to traverse to stream network nodes variable created to measure the difficulty of reaching this water resource.

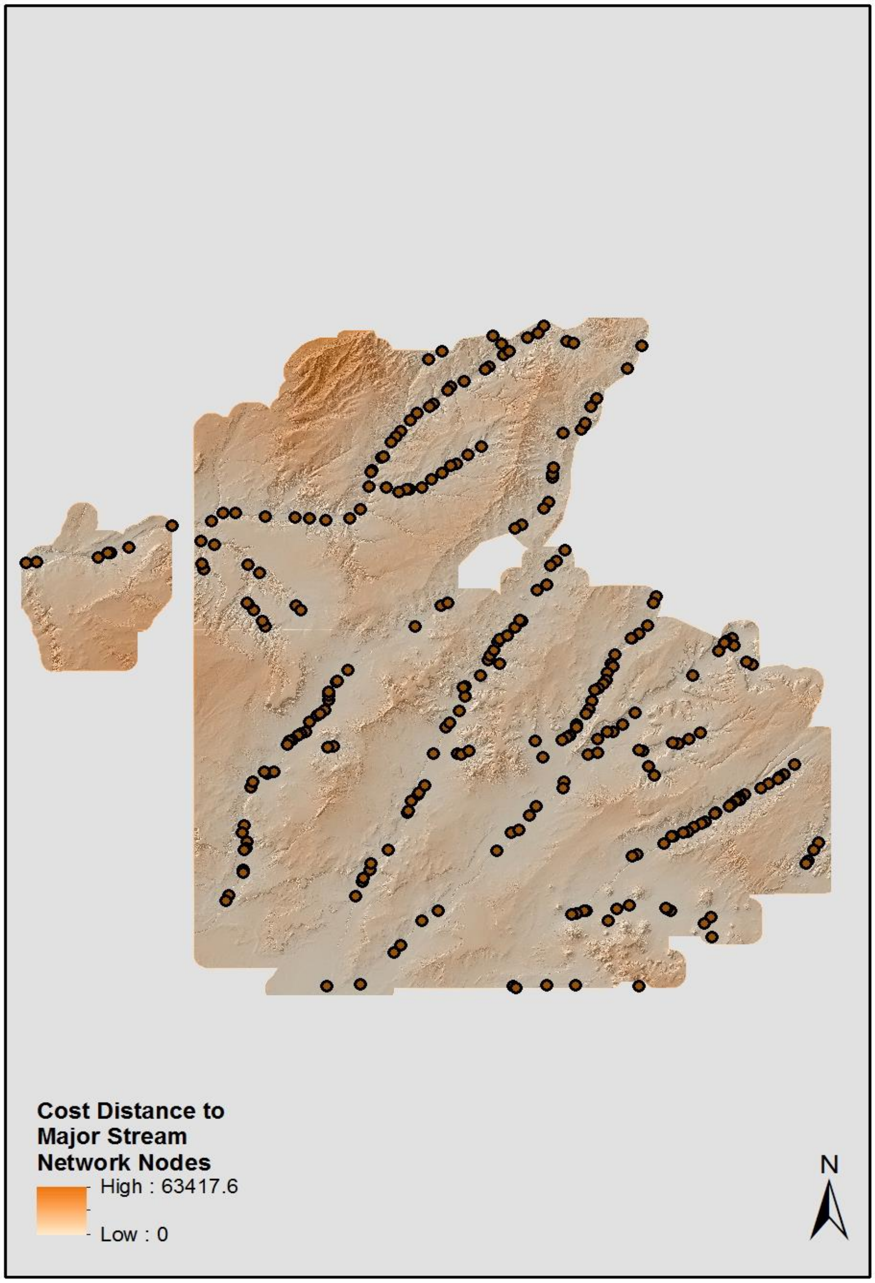
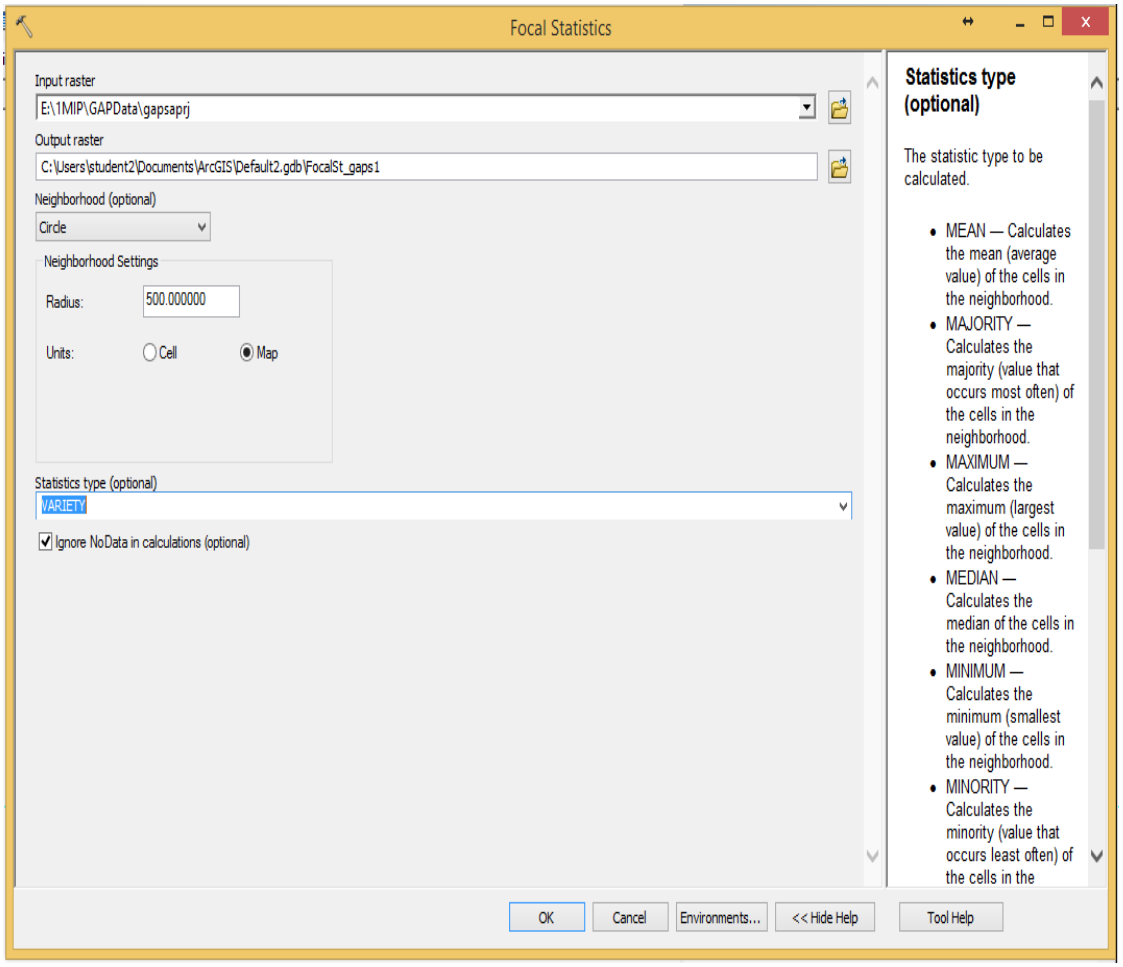


Figure 5-20: The cost to traverse to major stream network nodes variable created to determine the cost to traverse to the major stream network nodes.

Vegetation richness within 100- and 500-meter radii were the final variables developed for model building. Vegetation richness is the count of vegetation types found in a specified area (McGinley, 2011). It is believed that sites located near multiple vegetation types "also have greater access to a wider variety of resources than sites located in an area of uniform vegetation" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). The variable was created with the Focal Statistics geoprocessing tool by using a 100- and 500-meter radii and selecting the "Variety" statistics type [Figures 5-21, 5.22, and 5.23].


Focal Statistics

Input raster
E:\1MIP\GAPData\gapsaprj

Output raster
C:\Users\student2\Documents\ArcGIS\Default2.gdb\FocalSt_gaps1

Neighborhood (optional)
Circle

Neighborhood Settings
Radius: 500.000000
Units: ○ Cell ● Map

Statistics type (optional)
VARIETY

☑ Ignore NoData in calculations (optional)

Statistics type (optional)

The statistic type to be calculated.

- MEAN — Calculates the mean (average value) of the cells in the neighborhood.
- MAJORITY — Calculates the majority (value that occurs most often) of the cells in the neighborhood.
- MAXIMUM — Calculates the maximum (largest value) of the cells in the neighborhood.
- MEDIAN — Calculates the median of the cells in the neighborhood.
- MINIMUM — Calculates the minimum (smallest value) of the cells in the neighborhood.
- MINORITY — Calculates the minority (value that occurs least often) of the cells in the

Figure 5-21: The Focal Statistics geoprocessing tool for creating the vegetation richness variable.

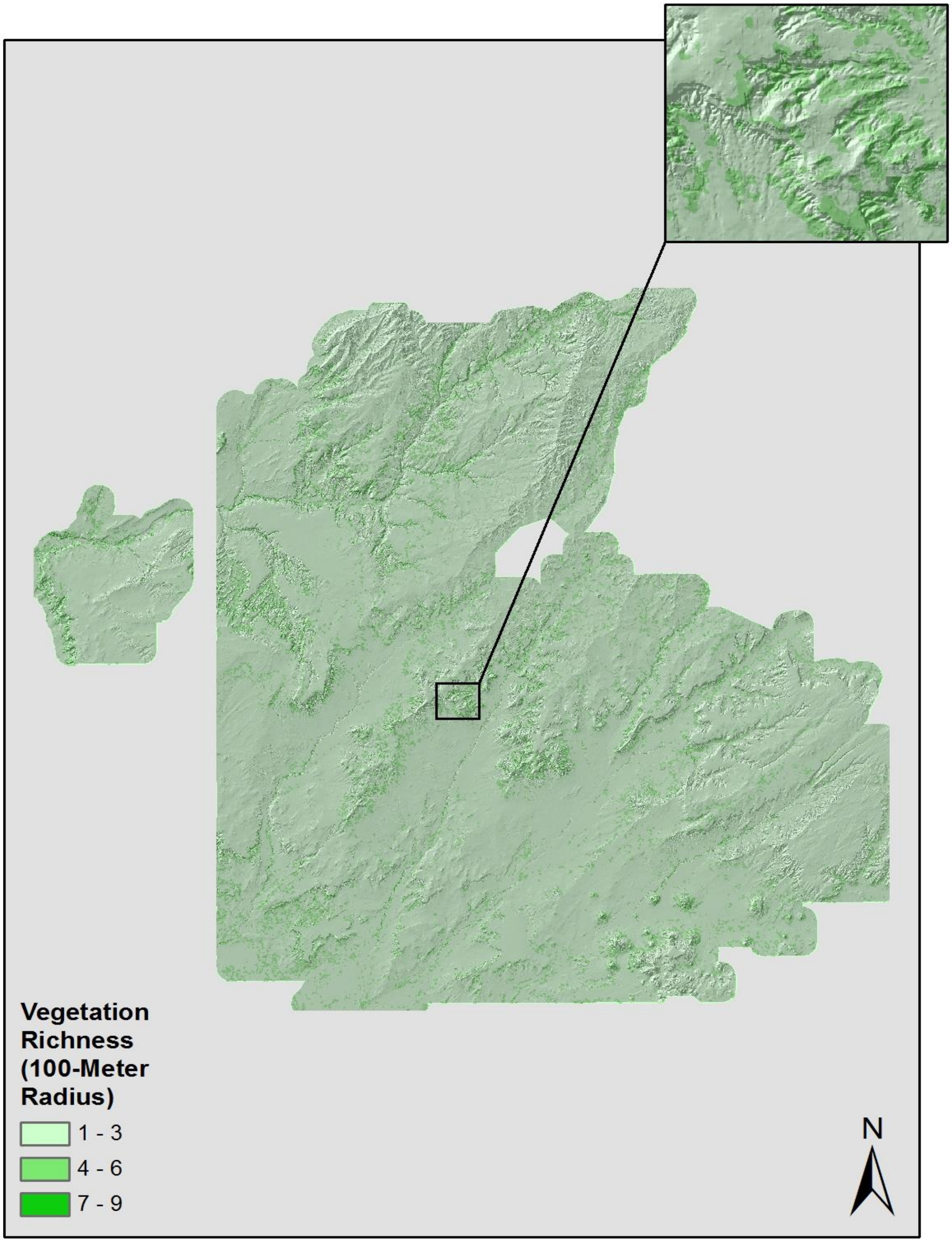


Figure 5-22: The vegetation richness variable created to determine the count of plant species within the specified 100-meter radius.

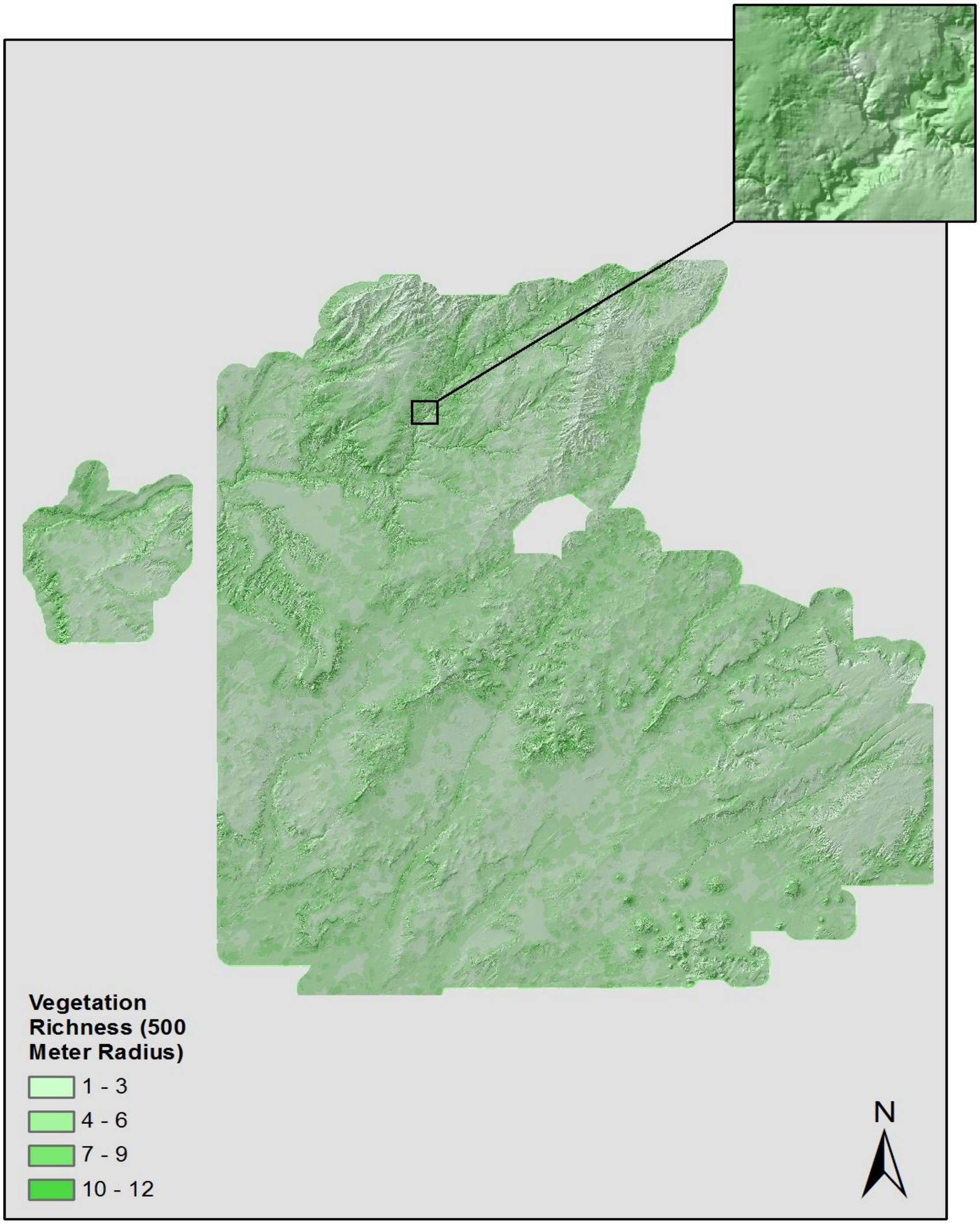
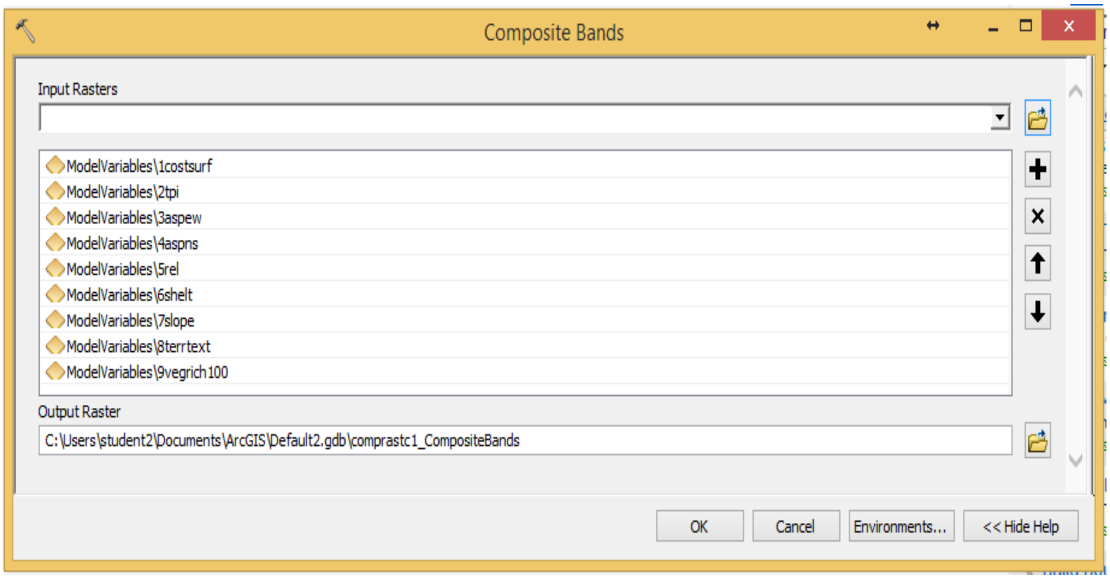


Figure 5-23: The vegetation richness variable created to determine the count of plant species with the specified 500-meter radius.

5.3 Variable Import

The variables then had to be further processed to import the values to SPSS. The Composite Bands geoprocessing tool was used to create a single composite raster of all 16 variables (Figure 5-24). This was done so that the raster variables would be properly aligned when the pixel values were extracted from ArcGIS 10.2 to a DBF file. The Site_Non-Site field was created for each archaeological site type feature class and non-site feature class. This field for the archaeological site type feature classes was populated with 1 to indicate site presence, and the same field for the non-site feature classes was populated with 0 to indicate site absence. Each archaeological site type feature class was then merged with the non-site feature class that had been created for it, creating three distinct feature classes. The Extract Multi Values to Points geoprocessing tool was utilized for each site type to extract the values from the newly created composite raster that was composed of all the independent variables. Once these values were extracted and saved in individual DBF files per site type, the logistic regression model was ready to be built in SPSS.



Input Rasters

ModelVariables\1costsurf
ModelVariables\2tpi
ModelVariables\3aspew
ModelVariables\4aspns
ModelVariables\5rel
ModelVariables\6shelt
ModelVariables\7slope
ModelVariables\8terrtext
ModelVariables\9vegrich100

Output Raster

C:\Users\student2\Documents\ArcGIS\Default2.gdb\comprastc1_CompositeBands

Figure 5-24: Composite Bands geoprocessing tool

5.4 Testing for Multicollinearity

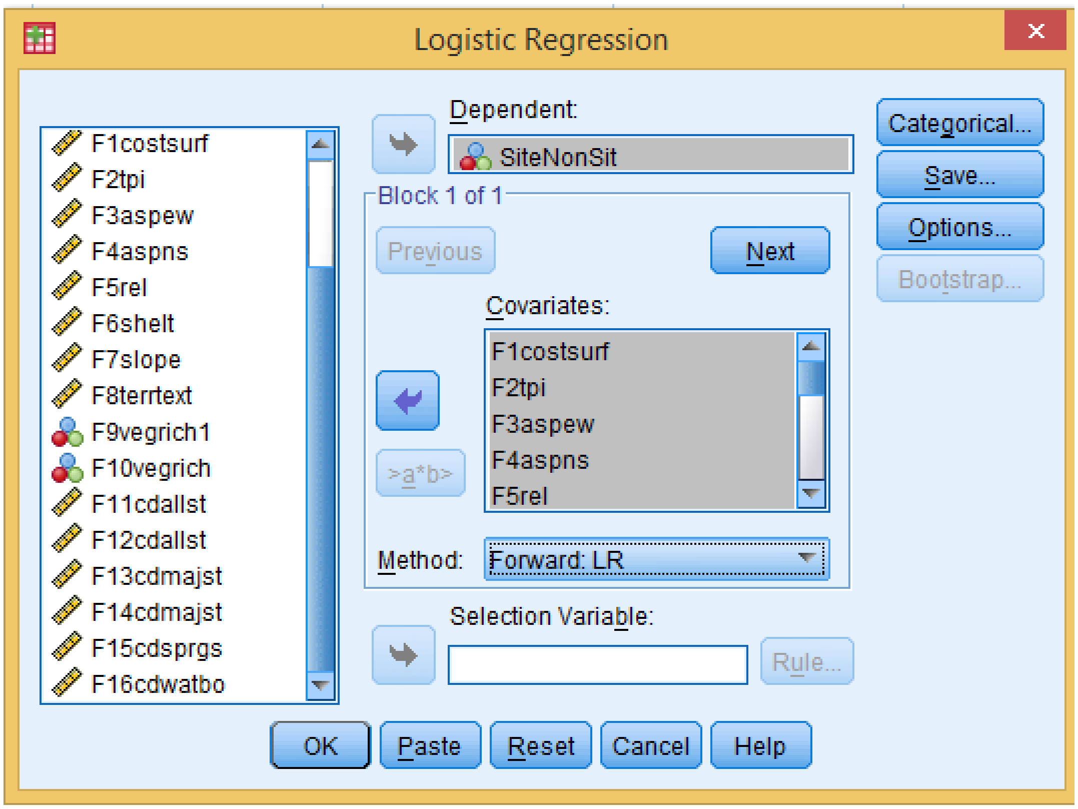
Once the variables were created and ready to be imported into SPSS, it was necessary to test for multicollinearity, or redundancy, among the independent variables. This was done by running the Ordinary Least Squares (OLS) regression geoprocessing tool in ArcGIS 10.2. The three feature classes that were previously created with the Extract Multi Values to Points that contain the independent variable values for each archaeological site type were input into the OLS geoprocessing tool. This tool created an output PDF report that provided the Variable Inflation Factor (VIF) metric for each independent variable in the model. According to ArcGIS Resource Center, if the VIF for each independent variable is above 7.5, the variable can be removed until only the independent variables in the model are below 7.5 (Esri, 2013). This was done for all the independent variables for each archaeological site type. This was taken into consideration when building models, but many models produced the best results when running the logistic regression model with all independent variables.

5.5 Logistic Regression Model Development in SPSS

Once the variable values had been extracted from ArcGIS 10.2 and saved in a DBF file, logistic regression models were run for each site type. This process took some time to determine what the best models for each site type. As discussed previously, the site /non-site type was always the dependent variable, which contained a value of either a 1 or 0 to indicate site presence or absence. The independent variables were the topographic, water resource, and vegetation variable values that had been imported in from ArcGIS. The Forward Likelihood Ratio and Forced Entry methodologies were utilized for the creation of these models to determine which independent variables have the highest correlation with the dependent variable, X [Figure 5-25]. Only variables with a p-value less than .10 were selected for building models. A significance level is the probability of committing a Type I error, or in other words, rejecting the null hypothesis that is, in fact true. It is common to use either a .10, .05, or .01 significance level for hypothesis testing (English, 2011). In this case, only variables that were significant at a 90% level of confidence ($p < .10$) were selected. The null hypothesis in this scenario is,

$$b_i = 0, i = l_i \cdots {}_{, k}$$

where b_i is the estimated coefficient for $x_{i,}$ and there are k independent variables. If the independent variable is found to be significant (i.e., $p < .10$), then the null hypothesis is rejected.



Dependent: SiteNonSit

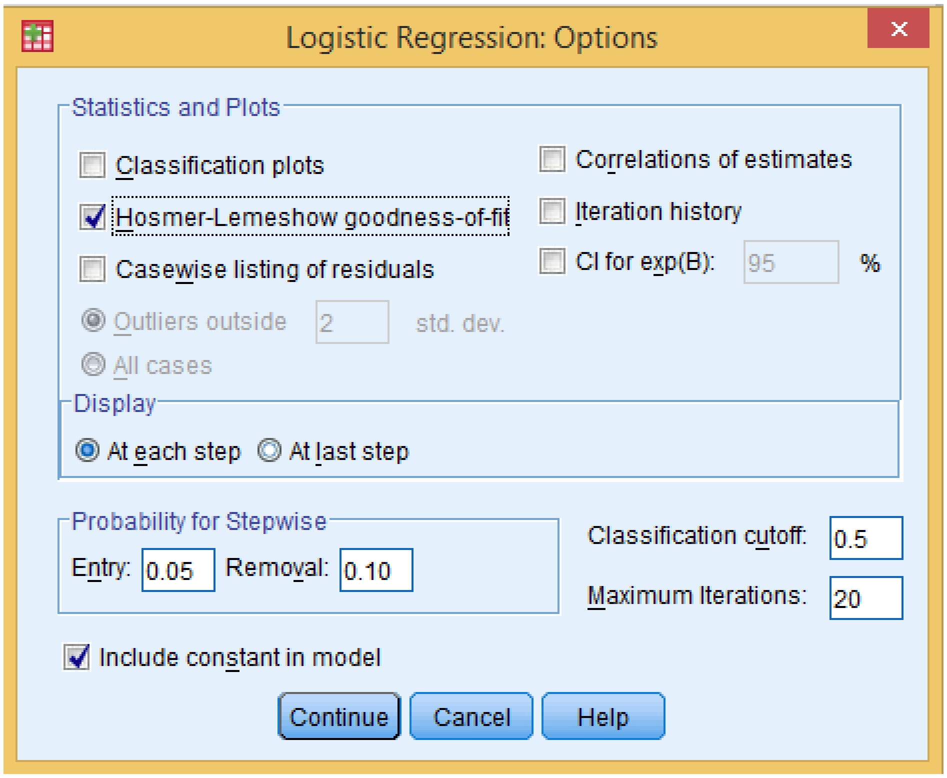
Covariates: F1costsurf, F2tpi, F3aspew, F4aspns, F5rel

Method: Forward: LR

Figure 5-25: Forward Likelihood Ratio methodology.

5.6 Model Validation Tests

The models had to pass the Hosmer-Lemeshow test and the omnibus tests of model coefficients to validate the models. For the omnibus tests of model coefficients, it is desirable for the models to have p-values less than .001. This tests the models' variables against a null model. This is a model with no variables and only constants. In all models, the significance levels were less than .001; therefore, the models were considered valid. Because the models passed this test, this indicated that at least one of the predictor variables is significantly related to the dependent variable (Apan et al., 2008). The Hosmer-Lemeshow goodness-of-fit-test was then utilized to test the overall fitness of the models [Figure 5-26]. A finding of non-significance indicates the model fits the data; therefore, it is desirable for the models' test statistics to have p-values above .05 (SPSS). All models passed this test.



Logistic Regression: Options

Statistics and Plots

☐ Classification plots

☑ Hosmer-Lemeshow goodness-of-fit

☐ Casewise listing of residuals

● Outliers outside 2 std. dev.

○ All cases

☐ Correlations of estimates

☐ Iteration history

☐ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

☑ Include constant in model

Figure 5-26: Hosmer-Lemeshow goodness-of-fit test

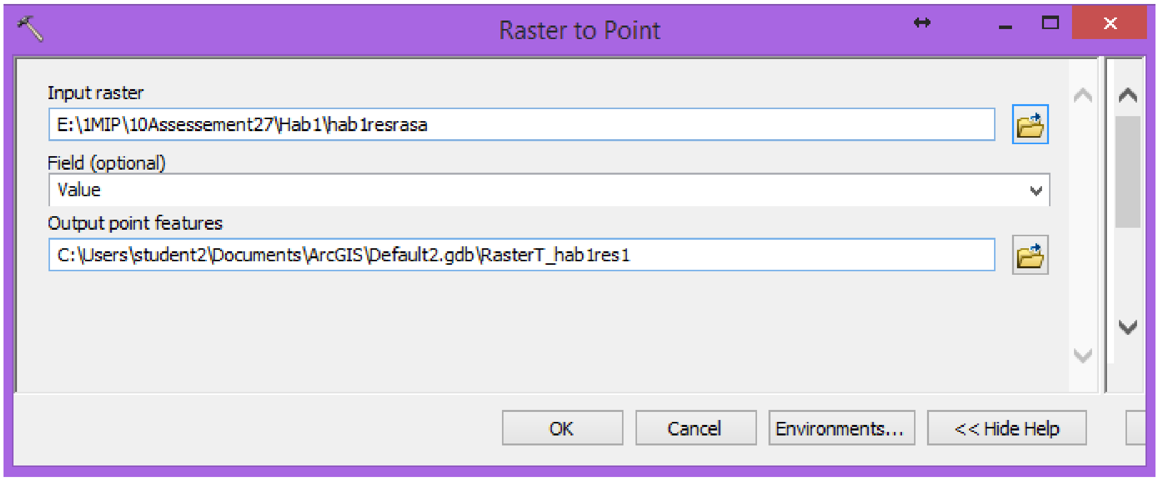
5.7 Model Assessment

Once the models were tested for validity and fitness, it was also pertinent to assess each model's level of accuracy. This was done by measuring the percentage of predicted values that were correctly classified for each model, by examining the classification tables in the SPSS outputs. The percentage correctly classified was calculated by determining the percentage of non-sites that were correctly classified and the percentage of sites that were correctly classified. This can be done by determining the number of hits that were correctly classified (Table 2). A hit in this scenario is either "1-1" or a "0-0", 1 for site and 0 for non-site (Denis, 2010). For non-sites (0), 79.2% were classified correctly. This was calculated by dividing the number of hits, 76, by the total of hits and non-hits, 96. This same formula was then applied to the sites (1). To determine the overall percentage, the total number hits (76, 81) were summed together (157) and divided by the total number of cases (188), which equaled 83.5%. It is also noteworthy to mention when a case is considered correctly classified. SPSS determined that .50 is the cutoff value; therefore, if a site is classified as .48, it considered to be incorrectly classified (Denis, 2010). By understanding the SPSS classification tables, one can quickly and efficiently determine which models are most accurate and worth considering further for research purposes.

Table 2. Classification table for assessing model accuracy.

Observed			Predicted		
			Non-Site	Site	Percentage Correct
			0	1	
Step 1	Non-Site	0	76	20	79.2
	Site	1	11	81	88.0
	Overall Percentage				83.5

The models were further assessed to determine what percentage of the Hopi Reservation had a probability of .5 or greater of finding the archaeological site in which the specified model was predicting. Because the reservation was so large in size, it was divided into four sections for the output of this geoprocessing tool to not exceed 2 gigabytes, the maximum allowed file size of a geoprocessing tool output. The reservation feature class was divided into four sections by using the Cut Polygons tool in the Editor toolbar and then utilizing the Extract by Mask geoprocessing tool to clip the predictive surface of the reservation into the four sections. The Raster to Point geoprocessing tool was then utilized to convert every pixel of the predictive surface to a point [Figure 5-27]. Each point contained the probability value of the associated pixel. The percentage of the reservation with probability of .5 or greater probability of finding the archaeological site in the specified model was then easily calculated.


Raster to Point

Input raster
E:\1MIP\10Assessement27\Hab1\hab1resrasa

Field (optional)
Value

Output point features
C:\Users\student2\Documents\ArcGIS\Default2.gdb\RasterT_hab1res1

Figure 5-27: The Raster to Point geoprocessing tool to raster pixels to points.

5.8 Predictive Raster Surface Creation

Once the models were built in SPSS, the predictive raster surfaces were created with the Raster Calculator in ArcGIS 10.2. For each archaeological site type, the following formula was implemented in the Raster Calculator to create predictive surfaces:

1 / (1 + (Exp(- (β + (b_1 * "Independent Raster Variable") + (b_2 * Independent Raster Variable") + (b_n * “Independent Raster Variable”)))))

in which β and b are the coefficients of the independent variables created in SPSS (Apan et al., 2008). This equation creates a predictive raster surface based on the coefficients, but it also creates an output probability surface ranging from 0 to 1, with 1 being the most likely to find an archaeological site. This equation makes the final outputs standardized and easy to interpret.

5.9 Summary

Chapter 5, Implementation, discussed the major phases of archaeological site predictive model development. This chapter was written to provide the audience with a detailed understanding of how the models were built. The chapter was divided into eight sections: Data Compilation and Creation Methods, Independent Variable Development, Variable Import, Testing for Multicollinearity, Logistic Regression Model Development in SPSS, Model Validation Tests, Predictive Raster Surface Creation, Model Assessment, and Summary. These topics are crucial to address in order for the client or another interested individual to replicate the process.

A large amount of time and effort were devoted to data preparation and independent variable development. Once the data was properly created and organized, a sufficient amount of time was devoted to building the models in SPSS and interpreting the results; although, the majority of the project work was devoted to the initial phases of the project, data preparation and independent variable development. It was beneficial to learn the importance of the initial organization and creation of the data. Good project planning and data organization were crucial for efficiency and project success.

Chapter 6 – Results and Analysis

To conclude the project, seven predictive models were produced for Rock Art, Habitation, and Scatter archaeological site types. Sixteen independent variables were created to be used in building the models. This chapter discusses model validity by addressing the outcomes of the omnibus tests of model coefficients and the Hosmer-Lemeshow goodness-of-fit test. This chapter also assesses model accuracy by examining the classification tables produced in the SPSS logistic regression output reports. Chapter 6 also reflects on what the final predictive raster surfaces communicate, what was successful in the project, and where there might have been errors. By addressing these areas, the audience will be able to see which models were most useful and which ones could be improved.

6.1 Model Validity

When the models were generated in SPSS, it was important to conduct the omnibus tests of model coefficients and the Hosmer-Lemeshow goodness-of-fit test to test for model validity. The omnibus tests of model coefficients are conducted to explain whether the predictor variables are "able to predict the dependent variable better than by chance alone" (Denis, 2010). If the results are statistically significant (p < .001), "the model does better than chance at predicting the dependent variable" (Denis, 2010). These tests indicate whether the model is worth investigating further. The Hosmer-Lemeshow goodness-of-fit test is a commonly used test that examines the overall fitness of the model. The idea behind this test is to see how well the model fits the data. In other words, the Hosmer-Lemeshow test examines whether "the model provides a better fit than a null model with no predictors, [only constants]…. If chi-square goodness-of-fit is not significant (p > .05), then the model has a good fit" (University of Strathclyde, n.d.).

Three logistic regression models were created for Rock Art. In the first model, the Forward Likelihood Ratio method was employed to determine the independent variables that were significant at a .10 significance level. These independent variables were east-west aspect, slope, and terrain texture. For this model, the omnibus test for model coefficients validated that at least one independent variable was able to predict the dependent variable with a p-value of < .001 (Table 3). From this, it could be surmised that at least one independent variable in the model could predict the dependent variable, site/non-site, better than by just chance alone. The next model, utilizing the Forced Entry methodology for all significant variables, also passed the omnibus tests of model coefficients (i.e. east-west aspect and all stream nodes). This is verified by the fact that the model test was significant with a p-value of < .001 (Table 3). It could then be surmised that at least one of the predictor variables can adequately predict the dependent variable. The final model for Rock Art also utilized the Forced Entry methodology for only the topographic and vegetation variables. The significant variables in this model were east-west aspect and terrain texture. The omnibus tests for model coefficients also validated that the predictor variables in the model adequately explained the dependent variable. This is verified by the fact that the model test with a p-value of < .001 (Table 3).

Table 3. Omnibus tests for model coefficients for Rock Art Models 1, 2, and 3.

Model 1		Chi Square	df	Sig.
Step 1	Step	142.552	1	<.001
	Block	142.552	1	<.001
	Model	142.552	1	<.001
Step 2	Step	11.812	1	.001
	Block	154.363	2	<.001
	Model	154.363	2	<.001
Step 3	Step	9.826	1	.002
	Block	164.189	3	<.001
	Model	164.189	3	<.001
Model 2		**Chi Square**	**df**	**Sig.**
Step 1	Step	172.285	16	<.001
	Block	172.285	16	<.001
	Model	172.285	16	<.001
Model 3		**Chi Square**	**df**	**Sig.**
Step 1	Step	165.848	10	<.001
	Block	165.848	10	<.001
	Model	165.848	10	<.001

The Hosmer-Lemeshow indicated positive results for the first Rock Art model. In this test for the first Rock Art model, the chi-square goodness-of-fit test results were not significant, with a p-value of .373 (Table 4). This suggested that the null hypothesis—the independent variables in the model adequately explain the dependent variable, site/non-site—is not rejected. Therefore, one could assert that the model adequately fits the data. This test also indicated positive results for the second Rock Art model. In this test for Rock Art, the chi-square goodness-of-fit test results was not significant, with a p-value of .572 (Table 4). This indicated the same results, that the null hypothesis—the independent variables in the model adequately explain the dependent variable, site/non-site—is not rejected. This test, again, indicated positive results for the third Rock Art model. In this test for Rock Art, the chi-square goodness-of-fit test results were not significant, with a p-value of .490 (Table 4). One can then assert that the model adequately fits the data.

Table 4. Hosmer-Lemeshow goodness-of-fit test for Rock Art Models 1, 2, and 3.

Model 1	**Chi-Square**	**df**	**Sig.**
Step 1	29.648	8	.490
Step 2	8.231	8	.411
Step 3	8.561	8	.373
Model 2	**Chi-Square**	**df**	**Sig**
Step 1	6.676	8	.572
Model 3	**Chi-Square**	**df**	**Sig.**
Step 1	7.445	8	.490

For the first Scatter site model, the omnibus tests for model coefficients utilizing the Forced Entry methodology for all variables verified that they adequately described the dependent variable. This is substantiated by the fact that omnibus tests of model coefficient tests were significant with a p-value of <.001 (Table 5). From this, it was surmised that the four significant variables (i.e. vegetation richness, shelter, cost, and cost to traverse to water bodies) used in the model can predict the dependent variable better than by just chance. The final Scatter site model utilizing the Forward Likelihood methodology also passed the omnibus tests for model coefficients with a p-value of <.001, indicating that at least one of the independent variables can predict the dependent variable better than by chance alone (Table 5). The significant variables at a .90 confidence level were vegetation richness with a 500-meter radius, slope, terrain texture, shelter, and major stream nodes.

Table 5. Omnibus tests for model coefficients for Scatter site models 1 and 2.

Model 1		**Chi Square**	**df**	**Sig.**
Step 1	Step	55.007	10	<.001
	Block	55.007	10	<.001
	Model	55.007	10	<.001
Model 2		**Chi Square**	**df**	**Sig.**
Step 1	Step	14.019	1	<.001
	Block	14.019	1	<.001
	Model	14.019	1	<.001
Step 2	Step	8.417	1	.004
	Block	22.436	2	<.001
	Model	22.436	2	<.001
Step 3	Step	6.674	1	.010
	Block	29.110	3	<.001
	Model	29.110	3	<.001
Step 4	Step	4.712	1	.030
	Block	33.822	4	<.001
	Model	33.822	4	<.001
Step 5	Step	6.413	1	.011
	Block	40.236	5	<.001
	Model	40.236	5	<.001

The Hosmer-Lemeshow goodness-of-fit test indicated positive results. In this test for Scatter, the chi-square goodness-of-fit test results were not significant, with a p-value of .121 (Table 7). The final Scatter site model also passed the Hosmer-Lemeshow test. In this test for Scatter, the chi-square goodness-of-fit test results were not significant, with a p-value of .544 (Table 7). These tests also confirms that the two Scatter site models adequately fit the data.

Table 6. Hosmer-Lemeshow goodness-of-fit test for Scatter site models 1 and 2.

Model 1	**Chi-Square**	**df**	**Sig.**
Step 1	12.755	8	.121
Model 2	**Chi-Square**	**df**	**Sig.**
Step 1	4.736	5	.449
Step 2	4.083	8	.850
Step 3	4.820	8	.777
Step 4	9.905	8	.272
Step 5	6.923	8	.544

The first Habitation site model utilized the Forced Entry methodology for all explanatory variables, but only included the independent variables that did not have a Variance Inflation Factor (VIF) higher than 7.5. Independent variables with a VIF above 7.5 are considered to be redundant and could possibly eliminated from the model (Esri, 2013). The significant variables in this model were vegetation richness with a 500-meter radius, slope, shelter, cost surface, cost to traverse to water bodies, and cost to traverse to all stream nodes. For this model, omnibus tests for model coefficients had positive results for with a p < .001, indicating that at least one of independent variables can predict the dependent variable better than by chance alone (Table 8). The Forced Entry methodology was also used for the final Habitation site model, but only included topographic and vegetation. The significant variables for this model were slope, shelter, cost surface, cost to traverse to water bodies, and cost to traverse to major stream nodes. This model also passed the omnibus test for model coefficients with a p-value < .001, indicating a valid model (Table 8).

Table 7. The omnibus tests for model coefficients for Habitation site models 1 and 2.

Model 1		Chi-Square	Df	Sig.
Step 1	Step	88.665	13	<.001
	Block	88.665	13	<.001
	Model	88.665	13	<.001
Model 2		**Chi-Square**	**Df**	**Sig.**
Step 1	Step	98.649	16	<.001
	Block	98.649	16	<.001
	Model	98.649	16	<.001

The first Habitation site model passed the Hosmer-Lemeshow goodness-of-fit test. In this test for the Habitation site model, the chi-square goodness-of-fit test results were not significant, with a p-value of .835 (Table 9). The final Habitation site model also passed the Hosmer-Lemeshow test. In this test for the Habitation site model, the chi-square goodness-of-fit test results were not significant, with a p-value of .110 (Table 9). By passing this test, it is confirmed the two Habitation models are valid models.

Table 8. The Hosmer-Lemeshow goodness-of fit test for Habitation site models 1 and 2.

Model 1	Chi-Square	df	Sig.
Step 1	4.241	8	.835
Model 2	**Chi-square**	**df**	**Sig.**
Step 1	13.066	8	.110

6.2 Model Assessment

For the first Rock Art model, which was created utilizing the Forward Likelihood methodology for all predictor variables, 89.4% of the sites and non-sites were correctly classified (Table 9). The significant variables at a 90 % confidence level were slope, east-west aspect, and terrain texture. The coefficients were .147, .029, and .064, respectively. This was one of the most accurate models based on this metric and also based on visual inspection [Figure 6-1]. The results from this model outputs explain that topography was the primary driver behind the model.

For the second Rock Art model which was created utilizing the Forced Entry methodology for all predictor variables, 92% of the sites and non-sites were correctly classified (Table 9). The significant variables in this model were only east-west aspect and cost to traverse to all stream nodes. The coefficients were .033 and .001, respectively. Although, this model had the highest percentage correctly classified, there were only two significant variables in the model and current Rock Art sites did not align very well with the predictive surface [Figure 6-2]. This should not be the primary indicator of what Rock Art model is most accurate, but these are good indicators as to which models are most worthy to further explore.

For the third Rock Art model which was created utilizing the Forced Entry methodology for only topographic and vegetation explanatory variables, 91% of the sites and non-sites were correctly classified (Table 9). The two significant variables in this model were only east-west aspect and terrain texture. The coefficients were .031 and .077, respectively. This model did have a very high percentage correctly classified and the current Rock Art sites appeared to align very well with where the predictive surface created from this model indicated a high likelihood of Rock Art sites [Figure 6-3]. This model was driven by only a few topographic variables, east-west aspect and terrain texture, but appeared to indicate a high likelihood of finding Rock Art sites in areas that appeared logical. These indicators can help determine which model is the most accurate, but further field investigation is recommended to truly determine which model is most accurate.

Table 9. Classification table for Rock Art logistic regression models 1, 2, and 3.

Model 1	**Predicted**		
Observed	**Non-Sites**	**Sites**	**Percentage Correct**
	0	**1**	
Step 1 Sites 1	9	83	90.2
Non-Sites 0	86	10	89.6
	Percentage		89.9
Step 2 Site 1	87	83	90.2
Non-Site 0	9	9	90.6
	Percentage		90.4
Step 3 Site 1	11	81	88.0
Non-Site 0	87	9	90.6
	Final Percentage		**89.4**
Model 2			
Step 1 Site 1	7	85	92.4
Non-Site 0	88	8	91.7
	Percentage		92.0
Model 3			
Site 1	8	84	91.3
Non-Site 0	87	9	90.6
	Final Percentage		**91.0**

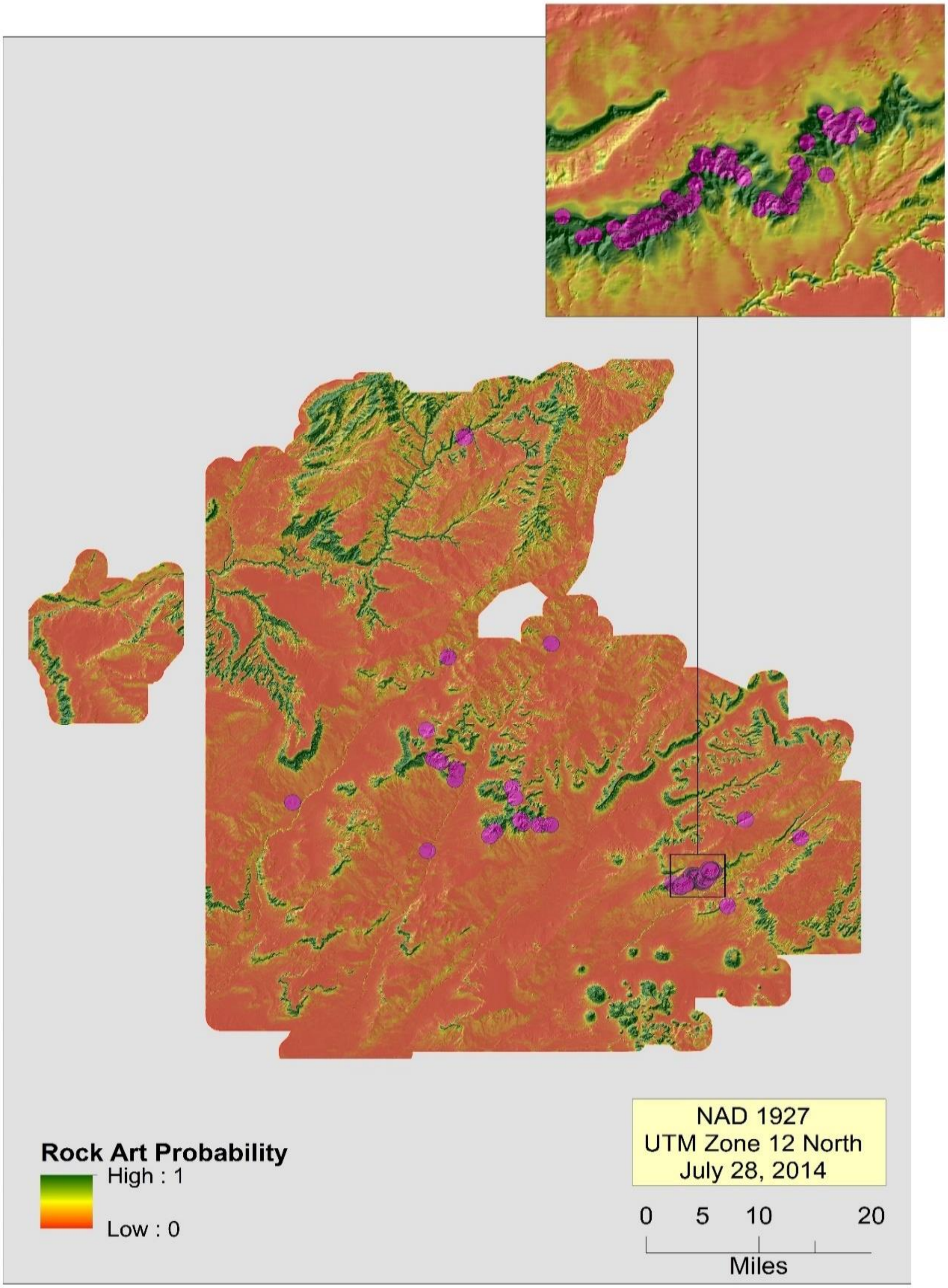


Figure 6-1: Distribution of current Rock Art sites on the predictive raster surface utilizing slope, east-west aspect, and terrain texture predictive variables.

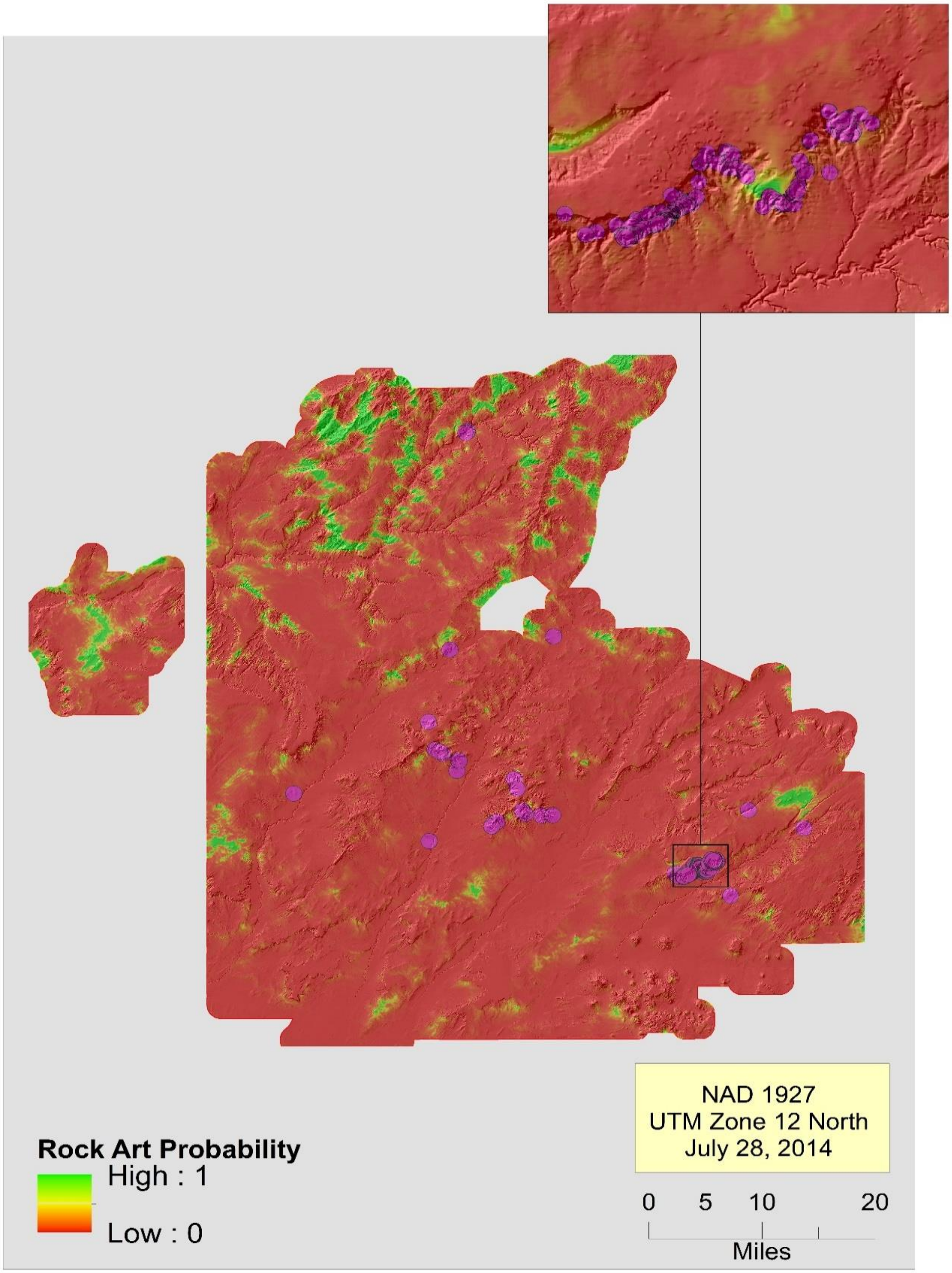


Figure 6-2: Distribution of current Rock Art sites on the predictive Rock Art surface utilizing east-west aspect and cost to traverse to all stream nodes predictive variables.

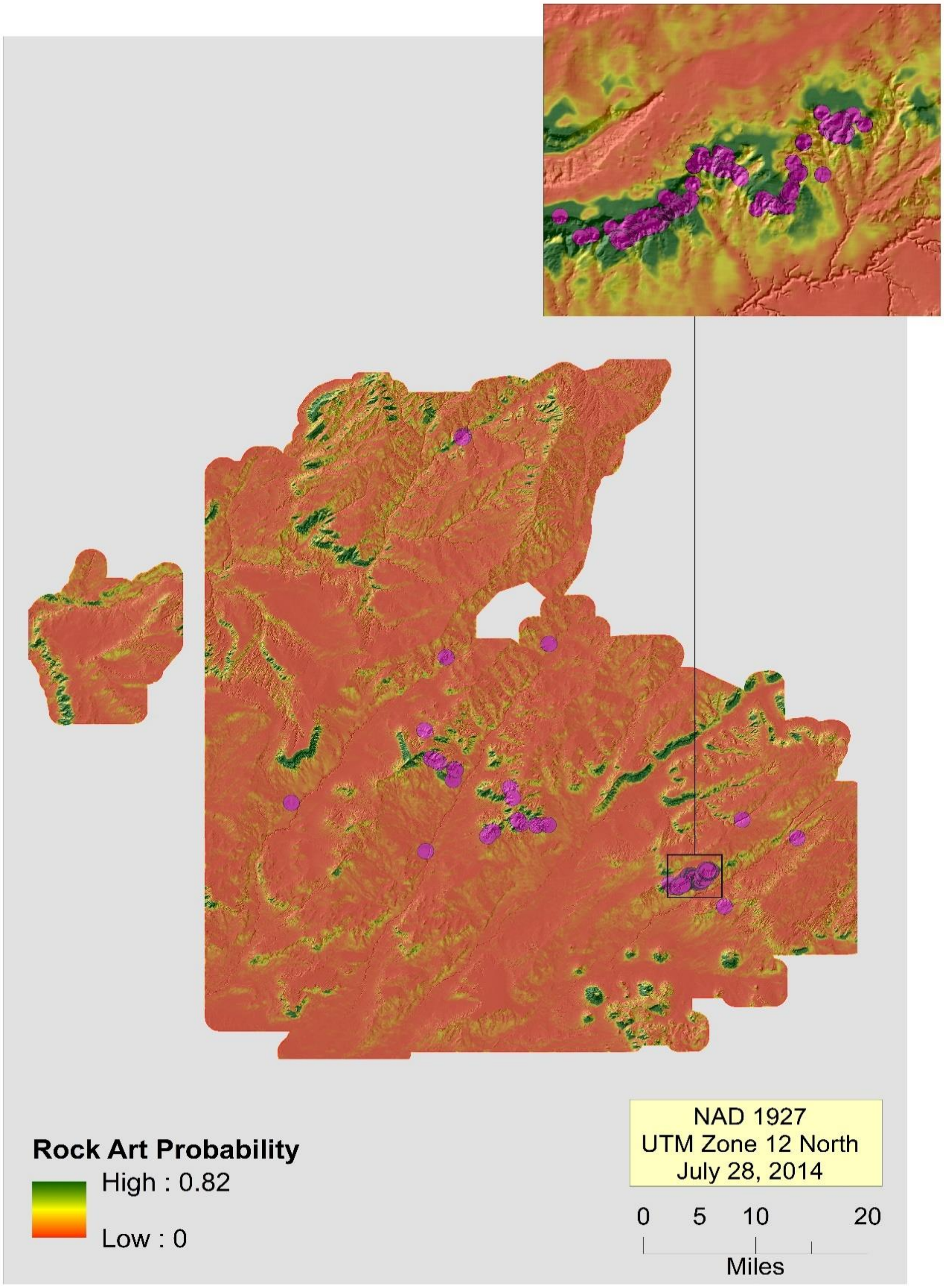


Figure 6-3: Distribution of current Rock Art sites on the predictive Rock Art surface utilizing east-west aspect and terrain texture explanatory variables.

For the first Scatter site model, which was created utilizing the Forced Entry methodology for all explanatory variables, 64.7% of the sites and non-sites were correctly classified (Table 10). The significant variables in this model were vegetation richness with a 500-meter radius, shelter, cost surface, and cost to traverse to water bodies. The coefficients were .199, -37.01, .46, and -.0000157, respectively. This model had the highest percent correctly classified, but was still relatively low compared to the Rock Art site models. The significant variables that contributed to the creation of this predictive surface came from all three variable categories, topography, vegetation, and water resources [Figure 6-4]. Although, this model scored the highest for Scatter sites, it appears that there could be further investigation into the Scatter site model to determine where it could be improved.

For the final Scatter site model which was created utilizing the Forward Likelihood methodology for all explanatory variables, 61.6% of the sites and non-sites were correctly classified (Table 10). The significant variables in this model were vegetation richness with a 500-meter radius, slope, terrain texture, shelter, and cost to traverse to major stream nodes. The coefficients were .224, -.051, .031, -33.318, and -.0000358, respectively. This model scored the lowest percent correctly classified for the Scatter site models. The significant variables that contributed to the creation of this predictive surface also came from all three variable categories, topography, vegetation, and water resources [Figure 6-5]. Like the previous Scatter site model, it appears that there could be further investigation into the Scatter site model to determine where it could be improved.

Table 10. Classification table for the Scatter site models 1 and 2.

Model 1	Predicted		
Observed	Non-Sites	Sites	Percentage Correct
	0	1	
Step 1 Sites 1	62	156	71.6
Non-Sites 0	119	88	57.5
	Final Percentage		64.7
Model 2			
Step 1 Sites 1	60	158	72.5
Non-Sites 0	93	114	44.9
	Percentage		59.1
Step 2 Site 1	84	134	61.5
Non-Site 0	114	93	55.1
	Percentage		58.4
Step 3 Site 1	82	136	62.4
Non-Site 0	115	92	55.6
	Percentage		59.1
Step 4 Site 1	65	159	70.2
Non-Site 0	115	92	55.6
	Percentage		63.1
Step 5 Site 1	76	142	65.1

Non-Site 0	120	87	58.0
	Final Percentage		**61.6**

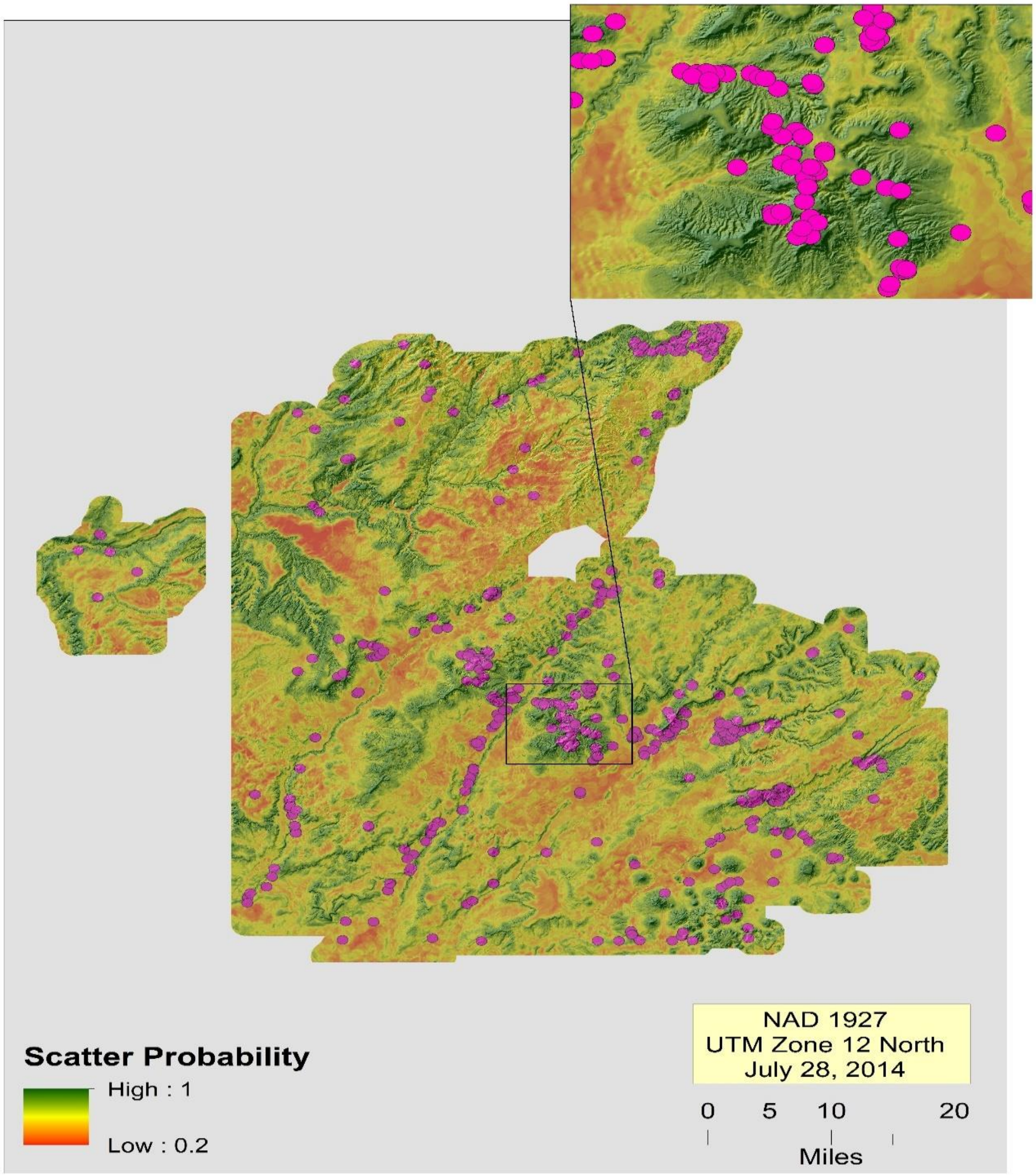


Figure 6-4: Distribution of current Scatter sites on the predictive surface utilizing vegetation richness at 500-meters, shelter, cost surface, and cost to traverse to water bodies explanatory variables.

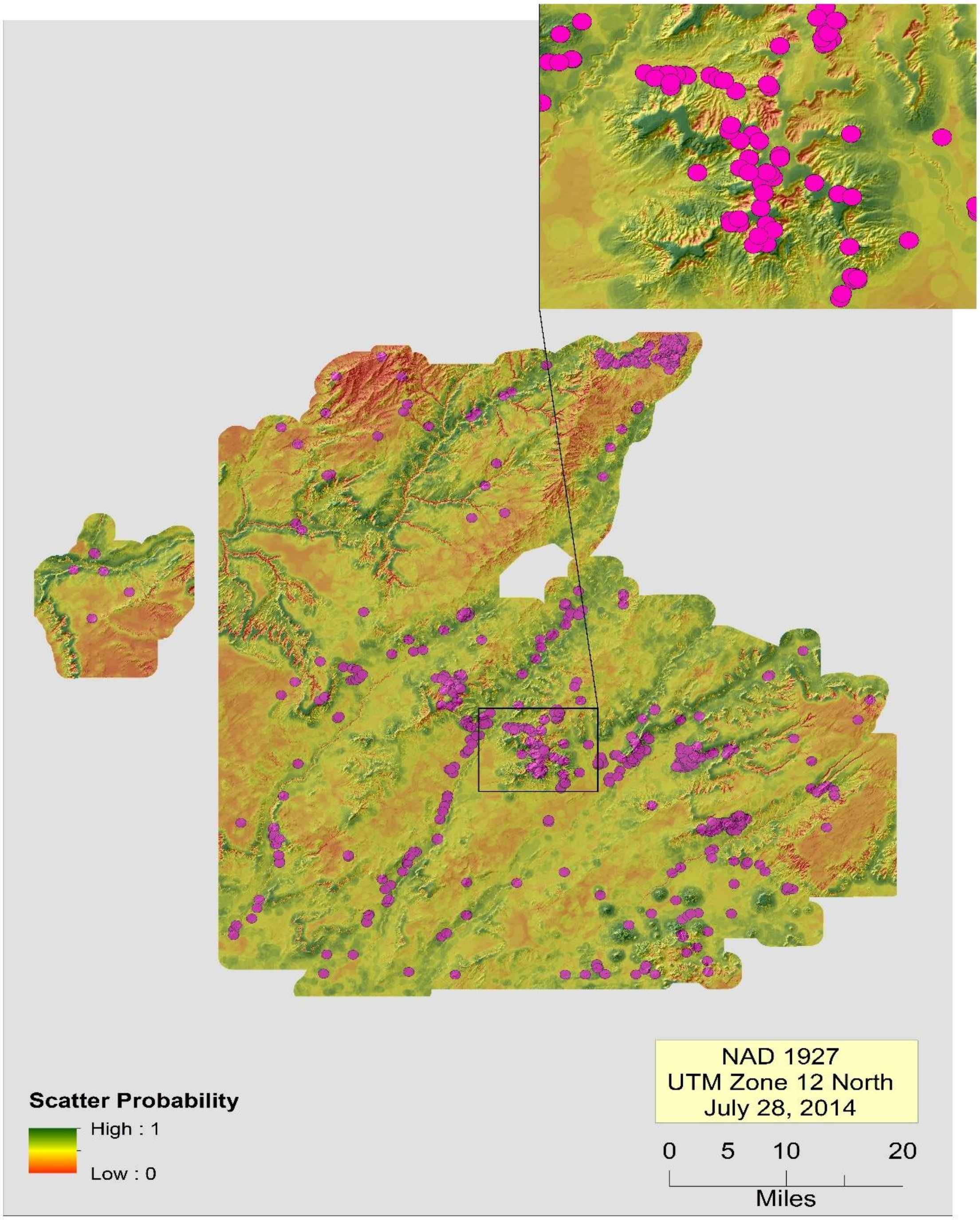


Figure 6-5: Distribution of current Scatter sites on the predictive surface utilizing vegetation richness with a 500-meter radius, slope, shelter, and cost distance to major stream nodes explanatory variables.

For the first Habitation site model which was created utilizing the Forced Entry methodology, 68.6% of the sites and non-sites were correctly classified (Table 12). The only variables not utilized in this model were the variables that had a Variance Inflation Factor above a 7.5, indicating autocorrelation. These variables were cost to traverse to major streams, cost to traverse to major stream nodes, and terrain texture. The significant variables in this model were vegetation richness with a 500-meter radius, slope, shelter, cost surface, cost to traverse to water bodies, and cost to traverse to all stream nodes. The coefficients were .131, -.112, -58.913, 1.029, -.0000224, and -.00001486, respectively. Although, this model indicated a relatively low percent correctly classified, it was interesting and noteworthy to observe that the model indicated high site likelihood on mesa tops [Figure 6-6]. This was most likely influenced by the negatively weighted shelter explanatory variable (-58.913). This indicator appears to be logical and correct due to the fact that many of the current habitation sites are located on top of mesas. Even though the percent correctly classified was relatively low, it is worthy to investigate this model further, possibly through fieldwork, to test how useful this model is and whether the significant explanatory variables truly indicate where settlements tended to be located.

For the final Habitation site model which was created utilizing the Forced Entry methodology, 68.6% of the sites and non-sites were also correctly classified (Table 23). The significant variables in this model were slope, shelter, cost surface, cost to traverse to water bodies, and cost to traverse to major stream nodes. The coefficients were -.114, -60.694, 1.07, -.0000206, and -.0000154, respectively. This model also scored a low percent correctly classified, but it was interesting and noteworthy to observe that the model indicated high likelihood near the major streams junctions. This indicator appeared to be logical and correct due to the fact that many of the current habitation sites are also located along major streams [Figure 6-6]. Like the previous model, even though the percent correctly classified was relatively low, it is worthy to investigate this model to test how useful this model is and whether the significant explanatory variables truly indicate where settlements tended to be located.

Table 11. Classification table for the Habitation site models 1 and 2.

Model 1	**Predicted**		
Observed	**Non-Sites**	**Sites**	**Percentage Correct**
	0	1	
Step 1 Sites 1	62	191	75.5
Non-Sites 0	128	84	60.4
	Final Percentage		**68.6**
Model 2			
Step 1 Sites 1	65	188	74.3
Non-Sites 0	131	81	61.9
	Final Percentage		**68.6**

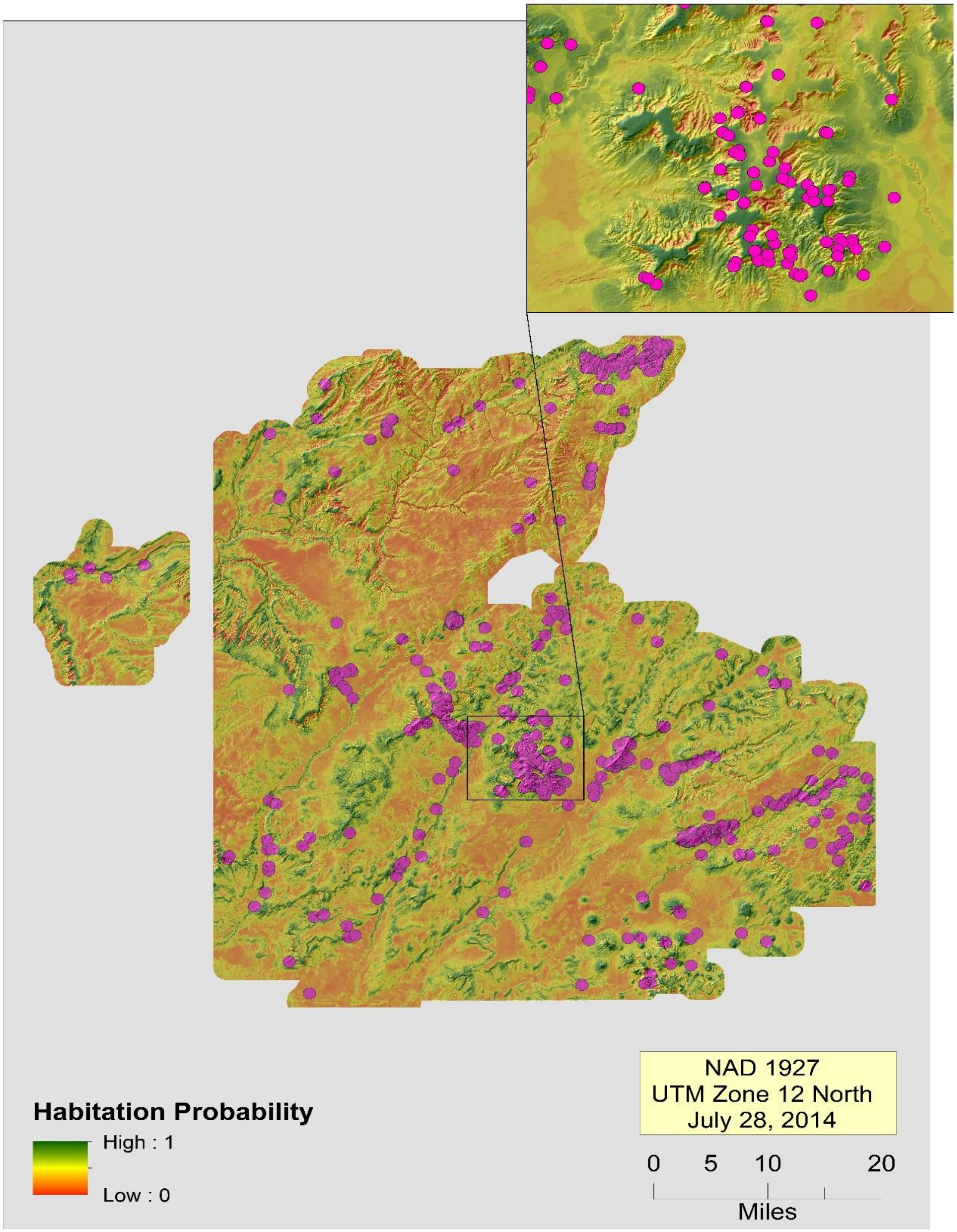


Figure 6-6: Distribution of current Habitation sites on the predictive surface utilizing vegetation richness with a 500-meter radius, slope, shelter, and cost distance to water bodies, and cost to traverse to all stream nodes explanatory variables.

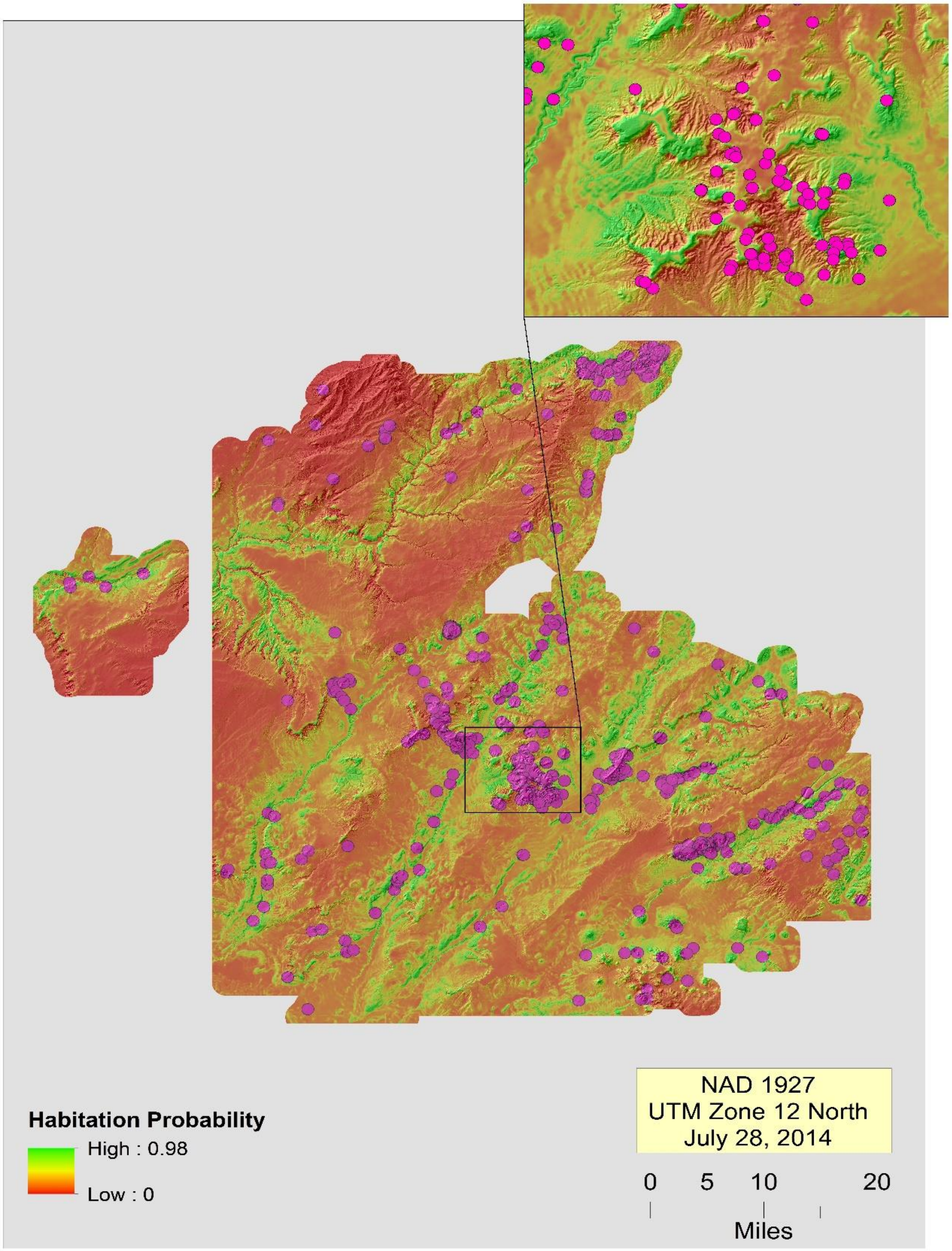


Figure 6-7: Distribution of current Habitation sites on the predictive surface utilizing slope, shelter, and cost surface, cost to traverse to water bodies, and cost distance to all stream nodes explanatory variables.

The final assessment of each model was to determine the percentage of each model's predictive surface that contained a probability of 0.5 or greater. This was done for all seven models. For the first Rock Art site model, 11.8 % of the reservation had a probability of 0.5 or greater to find a Rock Art site type. Rock Art models 1 and 2 contained 0.5 % and 0.02 % of the reservation, respectively. The average percentage for all three models was 4.1 % (Table 12).

Table 12. Percentage of Rock Art site models predictive surface that contained a probability of 0.5 >= of finding a Rock Art site type.

Rock Art Site Models	Percentage of Reservation with Probability Surface .5 >= of Finding Rock Art Sites
Model 1	11.8
Model 2	.5
Model 3	.0237
Average Percentage	**4.1**

For the first Habitation site model, 11.8 % of the reservation had a probability of 0.5 or greater to find a Rock Art site type. Rock Art models 1 and 2 contained 0.5 % and 0.02 % of the reservation, respectively. The average percentage for all three models was 4.1 % (Table 12).

Table 13. Percentage of Scatter site models predictive surface that contained a probability of 0.5 >= of finding a Scatter site type.

Scatter Site Models	Percentage of Reservation with Probability Surface 0.5 >= of Finding Scatter Sites
Model 1	89.6
Model 2	46.16
Average Percentage	**67.88**

For the habitation site models, the percentage of each model's predictive surface that contained a probability of 0.5 or greater was considerably higher. Habitation site models 1 and 2 contained a percentage of 27.16 % and 3.7 %, respectively. The average percentage for all three models was 15.47 % (Table 14).

Table 14. Percentage of Habitation site models predictive surface that contained a probability of 0.5 >= of finding a Scatter site type.

Habitation Site Models	Percentage of Reservation with Probability Surface .5 >= of Finding Habitation Sites
Model 1	27.16
Model 2	3.7
Average Percentage	**15.47**

6.3 Summary

In conclusion, each model produced remarkably different outputs, which was expected for each archaeological site type. The Rock Art archaeological site type models performed the best of the three models, statistically speaking. The Rock Art model that appeared to be the most successful had three topographic variables that were significant at a 90 % confidence level. This is intuitive because Rock Art should undoubtedly be found in areas of steeper slopes, which are areas where larger rocks are likely to be found (e.g., large boulders and canyon walls). The Scatter archaeological site type model is probably the model that needs to be refined the most. In discussion with the client, this archaeological site type possibly needs to be divided into more specific site types (Bernardini, personal communication, July 24, 2014). Many of the site type locations are located very close to the Habitation sites and followed the same clustering pattern [Figure 6-5, Figure 6-6]. This could be one of the many reasons this model appeared to not perform as well. Further investigation or reassessment of how this archaeological site type could be categorized into more specific site types is necessary. The Habitation site type models were the final models that were created. Although, the percent correctly classified was relatively low for these models, they both provided interesting results. The two models for Habitation received the exact same percent correctly classified scores (68.4%), but were dictated by slightly different variables that created different results. The first Habitation model addressed appeared to be more dictated by topographic variables. It was very interesting to note that some of the areas of highest likelihood were located on the mesa tops. This was probably due to the heavily negatively weighted shelter variable, which is logical because mesa tops provide very little shelter from the environmental elements. Conversely, the second Habitation model indicated that the areas with the highest likelihood of site location were associated with water resources. Many of the areas with the highest likelihood were in close proximity to the major streams or their junctions. These were two interesting model outputs and warrant further investigation of their validity and accuracy.

Chapter 7 – Conclusions and Future Work

The central goal of the project was to develop predictive models for the Rock Art, Scatter, and Habitation archaeological site types. This was accomplished by creating 16 independent raster variables in ArcGIS 10.2. There were three categories under which these variables were developed: topography, water resources, and vegetation. These independent variables were created to assist in determining the possible locations of the three archaeological site types. The dependent variables and sites/non-sites were also developed in ArcGIS 10.2 in order to run the logistic regression models in SPSS.

Once the independent and dependent variables were created in ArcGIS, logistic regression models were built in SPSS using various methods to select the variables that were significant at a 90% confidence level. Three models were then selected once they passed the Hosmer-Lemeshow goodness-of-fit test and the omnibus tests of model coefficients. The coefficients and constants created in the selected models were then input into the final algorithm that was used in the ArcGIS 10.2 Raster Calculator.

By doing this, the predictive surfaces for the archaeological site types were created. All models were assessed for accuracy by examining the classification tables in SPSS, which assessed the percentage of non-sites that were correctly classified and the percentage of sites that were correctly classified. A point was considered to be correctly classified if it was a site above .5 or a non-site below .5 (Denis, 2010).

All models indicated a level of accurate predictive capacity, but there are components of the model building process that could be improved. The following section addresses the following areas that might improve model success: soils variable development, historical-period resources variable development, and the Random Forests statistical approach.

7.1 Soils Variable

It is quite possible that introducing soils variables would improve the model. Multiple soils variables were utilized by SRI in building their predictive models. SRI staff observed that many sites were located on the edges of soil-mapping units (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Because of this, they used a combination of soils data from the Soil Survey Geographic Database (SSURGO) and State Soil Geographic Database (STATSGO). This might have to be done at the state level because not all soils in every state have been mapped at the same level of detail. Some of the variables SRI developed were cost to traverse to soil-texture boundary, standard deviation to soil-texture boundary, standard deviation in soil-texture index, and range in organic-matter content (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Other additional variables SRI created included the following:

- Available water capacity (inches of water per inch of soil profile)
- Bulk density (grams per cubic centimeter)
- Calcareousness (percent calcium carbonate)
- Cation-exchange capacity (molar equivalent per 100 g)
- Electrical conductivity (decisiemens per meter)

- Organic matter (percent organic matter by mass)

These were the majority of the soil variables that SRI considered when building their predictive models. In discussion with one of the principal investigators of the SRI project, Michael Heilen communicated that soils could most definitely be worth investigating for improved model performance (Heilen, personal communication, March 5, 2014).

7.2 Historical-Period Resources Variables

Historical-period resources are another variable that could be considered. These include transportation routes, such as trails, wagon roads, and railroads (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). These could have been important in predictive modeling "because much of the US West was dependent on the use of transportation for exploration and population migration as well as the redistribution for goods and materials" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). Because of this, it could be useful to create variables for cost to traverse to historical-period resources, cost to traverse to historical-period transportation routes, and cost-distance to [water] tanks. These data can be found in the USGS Geographic Names Information System database.

7.3 Random Forests Statistical Approach

In this study, the logistic regression analysis was utilized in building predictive models. In the other study, SRI used a relatively new statistical method called Random Forests. According to Heilen et al. (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012), "this approach samples both cases [dependent variable] and model [independent] variables during the course of the model development and tests model performance with samples not used to train the model." This points out a limitation of the logistic regression analysis and might be worthy of considering if someone is interested in improving model performance. It was also mentioned that the Random Forests method is "robust in overfitting from intercorrelations from variables" (Heilen, Leckman, Byrd, Homburg, & Heckman, 2012). This is another very important consideration because many of the independent variables were created from a DEM, which may have caused considerable intercorrelation among variables. These were a few of the reasons why the Random Forests statistical method to modeling might possibly be a better approach.

7.4 Summary

In conclusion, there are always strategies to take into consideration to improve model performance. It has been noted that the development of soils and historical resources variables could improve model performance. The Random Forests statistical approach was also proposed as another overarching strategy that may improve performance.

In the model, topographic variables were deemed to be the most important in building models. Vegetation richness with a 500-meter radius was also of primary importance in building the Scatter site model. In conclusion, it was found that the Rock Art and Habitation models were the most successful models and were of most use to the client.

Works Cited

Apan, A., Wells, N., Reardon-Smith, K. R., McDougall, K., & Basnet, B. (2008, July). *Predictive Mapping Of Blackberry In the Condamine Catchment Using Logistic Regression and Spatial Analysis.* Retrieved from www.usq.edu.au: http://eprints.usq.edu.au/4812/1/Apan_Wells_Reardon-Smith_Richardson_McDougall_Basnet.pdf

Cooley, S. (2014). *GIS 4 Geomorphology*. Retrieved from http://gis4geomorphology.com/: gis4geomorphology.com/roughness-topographic-position/

Denis, D. (2010). *Binary Logistic Regression Using SPSS.* Retrieved from University of Montana Department of Psychology: http://psychweb.psy.umt.edu/denis/datadecision/binary_logistic_spss/index.html

English, R. (2011, July 29). *Chapter 13 - Hypothesis Testing*. Retrieved from San Diego State University: Rohan Academic Computing: http://www-rohan.sdsu.edu/~renglish/470/notes/chapt13/chapter13.htm

Esri. (2011). *How Principal Component Analysis Works*. Retrieved from ArcGIS Resource Center: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Principal_Components_works/009z000000qm000000/

Esri. (2012). *Principal Components*. Retrieved from Principal Components (Spatial Analyst): http://resources.arcgis.com/en/help/main/10.1/index.html#//009z000000pq000000

Esri. (2013). *Interpreting OLS Results*. Retrieved from ArcGIS Resources: http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000030000000

Fattah, M. A. (2010). *Multicollinearity*. Retrieved from Central Michigan University: College of Humanities, Social, and Behavioral Sciences: http://www.chsbs. cmich.edu/fattah/courses/empirical/multicollinearity.html

Field, A. (2005). Discovering Statistics Using SPSS. London, England.

Fish, P. R. (2013, September 18). *Archaeological Sites*. Retrieved from Learning Center of the American Southwest: http://www.southwestlearning.org/topics/ archeological-sites

Goodchild, M. (1992). Geographic Information Systems Science. *International Journal of Geographic Information Systems*, 33.

Heilen, M., Leckman, P. O., Byrd, A., Homburg, J. A., & Heckman, R. A. (2012). *Archaeological Sensitivity Modeling in Southern New Mexico: Automated Tools and Models for Planning and Management.* Albuquerque: Statistical Research, Inc.

Hosmer, D., & Lemeshow, S. (2013). Applied Logistic Regression. Hoboken, New Jersey, United States: John Wiley & Sons.

Kohler, T., & Parker, S. (1986). Predictive Models for Archaeological Resource Locations. *Advances in Archaeological Methodology and Theory, Vol. 9*, 397-452.

Kvamme, K. (1988, December). *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling.* Denver:

U.S. Government Printing Office. Retrieved from Quantifying the Present and Predicting the Past: http://www.blm.gov/pgdata/etc/medialib/blm/wo/Planning_and_Renewable_Resources/coop_agencies/cr_publications.Par.58402.File.dat/quantifying_the_present.pdf

Leathwick, J. (2000, November). *Predictive Models of Archaeological Site Distributions in New Zealand.* Wellington: New Zealand Department of Conservation. Retrieved from Department of Conservation: Te Papa Atawhai: http://www.doc.govt.nz/documents/science-and-technical/ir181.pdf

McGinley, M. (2011, September 10). *Species Richness*. Retrieved from The Encyclopedia of Earth: http://www.eoearth.org/view/article/156216/

NetMBA. (2010). *Scatter Plot*. Retrieved from NetMBA: http://www.netmba.com/statistics/plot/scatter/

Sonoma State University. (2008). *Cultural Resource Management*. Retrieved from www.sonoma.edu: http://www.sonoma.edu/asc/aboutus/crm.htm

SPSS. (n.d.). Help Topics.

Strahler, S. M. (1957). Quantitative Analysis of Watershed Geomorphology. Burlington, Vermont, United States.

U.S. EPA. (n.d.). *National Hydrography Dataset Plus*. Retrieved from NHD Plus Version 2: http://www.horizon-systems.com/nhdplus/NHDPlusV1_CO.php

University of Strathclyde. (n.d.). *Goodness of Fit Measures*. Retrieved from University of Strathclyde: Humanities and Social Sciences: http://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/goodnessoffitmeasures/

USGS. (2011, August 3). *National Gap Analysis Program (GAP) | Land Cover Data Portal*. Retrieved from National Gap Analysis Program (GAP) -- Core Science Analytics and Synthesis: http://gapanalysis.usgs.gov/gaplandcover/

USGS. (2014, April 15). *National Elevation Dataset*. Retrieved from USGS: Science a Changing the World: http://ned.usgs.gov/

Weiss, A. D. (2014, June 29). *Topographic Position and Landforms Analysis.* Retrieved from Jenness Enterprises: http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf

Winemiller, D. T. (2014). GIS Reveals Basis for Ancient Settlement Location. *ArcNews*. Redlands, California, United States: Esri.