

University of Redlands

InSPIRe @ Redlands

MS GIS Program Major Individual Projects

Theses, Dissertations, and Honors Projects

12-2014

Predicting Archaeological Sensitivity of Puerto Rico using GIS

Loderay Isabel María Bracero-Marrero
University of Redlands

Follow this and additional works at: https://inspire.redlands.edu/gis_gradproj



Part of the [Geographic Information Sciences Commons](#), and the [Remote Sensing Commons](#)

Recommended Citation

Bracero-Marrero, L. I. (2014). *Predicting Archaeological Sensitivity of Puerto Rico using GIS* (Master's thesis, University of Redlands). Retrieved from https://inspire.redlands.edu/gis_gradproj/227



This work is licensed under a [Creative Commons Attribution 4.0 License](#).

This material may be protected by copyright law (Title 17 U.S. Code).

This Thesis is brought to you for free and open access by the Theses, Dissertations, and Honors Projects at InSPIRe @ Redlands. It has been accepted for inclusion in MS GIS Program Major Individual Projects by an authorized administrator of InSPIRe @ Redlands. For more information, please contact inspire@redlands.edu.

University of Redlands

Predicting Archaeological Sensitivity of Puerto Rico using GIS

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

by

Loderay I.M. Bracero-Marrero

Ruijin Ma, Ph.D., Committee Chair

Fang Ren, Ph.D.

December 2014

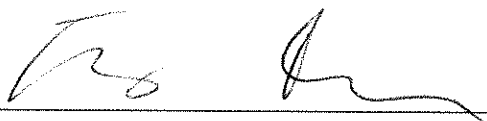
Predicting Archaeological Sensitivity of Puerto Rico using GIS

Copyright © 2014

by

Loderay I.M. Bracero-Marrero

The report of Loderay I.M. Bracero-Marrero is approved



Ren Fang, Ph.D.



Ruijin Ma, Ph.D., Committee Chair

December 2014

Acknowledgements

I want to thank Life and all His manifestations. The dialectal materialism could not go wrong. I want to thank my family, Jorge and Carola, for being there regardless of my absence and giving me their support with letters, gifts and coffee from Lares, Puerto Rico. In addition, I want to thank my mom (Isabel) and dad (Billy) for raising me and my other four siblings in life. Thank you for your support back in Puerto Rico.

I want to thank my advisor Ruijin for his support and patience around this year. He was a great advisor in all senses, personal and professionally. He inspired me to elevate my knowledge and become a better academic. I also want to thank Fang for all her support with the statistical topics that I dealt within this project. Lastly, I want to thank Debbie for being the maternal representation for us here during this year.

I want to thank my friends from Cohort 24 & 25. You know who you are. Debbie always said that there will be many changes (circles) in the Cohort, and she was completely right. Without the support and patience from my friends, this year experience would have been another story. It was a great comfort to have friends to have fun and study at the same time. In my opinion, this is the definition of a good friend. My hope is that life will always reunite us in different parts of the world and different situations. I want to thank my good friend of life, my conspiratorial friend Joseph for all the experience we had together.

This year I have been supported in both personal and professional life, especially to continue my doctorate studies. In this regard, I want to thank my forever advisor Reniel Rodríguez-Ramos for his support and believing in this project. This project could not have been possible without the support from Laura Del Olmo Frese. Thank you for your trust to start developing this project.

“The real education of the masses can never be separated from their independent political, and especially revolutionary, struggle. Only struggle educates the exploited class. Only struggle discloses to it the magnitude of its own power, widens its horizon, enhances its abilities, clarifies its mind, forges its will.”

-Vladimir Lenin, 1917, Lecture on the 1905 Revolution

Abstract

Predicting Archaeological Sensitivity of Puerto Rico using GIS

by

Loderay I.M. Bracero Marrero

The knowledge of Puerto Rican archaeological heritage has increased in different perspectives and approaches because of influences of several archaeological theories. The Puerto Rican archaeology scientific transformation applying geographic information systems (GIS) has been conducted but not with the full potential. This project used statistical and spatial analysis to study the archaeological site patterns in Puerto Rico. Using a logistic regression model (LRM), the project developed a predictive model for archaeological sites in Puerto Rico. The predictive model describes high, medium, and low likelihood areas where archaeological sites could be found. Tools and techniques such as Model Builder, photogrammetry analysis, and different sampling methods were implemented to accomplish the goal of this project.

Table of Contents

Chapter 1 – Introduction	1
1.1 Client.....	2
1.2 Problem Statement	2
1.3 Proposed Solution	2
1.3.1 Goals and Objectives	3
1.3.2 Scope.....	3
1.3.3 Methods.....	4
1.4 Audience	5
1.5 Overview of the Rest of This Report	5
Chapter 2 – Background and Literature Review	7
2.1 GIS in Archaeology	7
2.2 Predictive Models for Archaeology	8
2.2.1 Inductive vs. Deductive Methods	8
2.2.2 Environmental Determinism.....	9
2.2.3 Statistical Regression Methods.....	9
2.2.4 Limitations of PMA	10
2.2.5 Case Studies	11
2.3 Summary	12
Chapter 3 – Systems Analysis and Design.....	13
3.1 Problem Statement	13
3.2 Requirements Analysis	13
3.3 System Design	14
3.4 Project Plan	16
3.5 Summary	18
Chapter 4 – Database Design.....	19
4.1 Conceptual Data Model	19
4.2 Logical Data Model	19
4.3 Data Sources	20
4.4 Data Scrubbing and Loading	22
4.5 Summary	22
Chapter 5 – Implementation.....	23
5.1 Dependent Variables	23
5.2 Independent Variables	24
5.2.1 Topographic Variables.....	24
5.2.2 Soils and Land Cover Factors	29
5.2.3 Social Factors.....	31
5.3 Re-classifying Categorical Data	34
5.4 Associating Independent Variables to the Dependent Variable	35
5.5 Identifying Significant Independent Variables	36
5.5.1 Topographic Variables.....	36
5.5.2 Soil and Land Cover Variables	37
5.5.3 Social Variables	39

5.5.4	Logistic Regression Modeling	39
5.6	Probability Surface Creation.....	40
5.7	Summary	41
Chapter 6	– Results and Analysis.....	43
6.1	Logistic Regression Model Fitness.....	43
6.2	Probability Surface.....	44
6.3	Model Validation	46
6.4	Summary	47
Chapter 7	– Conclusions and Future Work.....	49
Works Cited	51
Appendix A.	Sample Code SAS.....	53

Table of Figures

Figure 1.1: Puerto Rico in Relation to the Caribbean Region	1
Figure 1.2: Study Area.....	4
Figure 3.1: System Design.....	15
Figure 3.2: Detailed Project Design.....	16
Figure 4.1: Conceptual Model	19
Figure 4.2: Logical Data Model.....	20
Figure 4.3: Example of Data Stored in the GBD	20
Figure 4.4: USDA Land Cover Data.....	21
Figure 4.5: Data Received from the Client.....	21
Figure 5.1: Site Samples and No-Site Samples	23
Figure 5.2: Topographic Variables	24
Figure 5.3: Elevation Surface	25
Figure 5.4: Slope Surface.....	26
Figure 5.5: Relief Surface	27
Figure 5.6: Suitability for Shelters.....	28
Figure 5.7: Aspect in Relationship to the South	29
Figure 5.8: Soils Drainage and Farmlands.....	30
Figure 5.9: Land Cover Data	31
Figure 5.10: Cost Surface	32
Figure 5.11: Cost Distance to Rivers and Creeks	33
Figure 5.12: Cost Distance to Coastlines.....	34
Figure 5.13: Scree Plot PCA Components.....	37
Figure 5.14: Frequency Charts of Farmlands Categories	38
Figure 5.15: Frequency Charts of Drainage Categories.....	38
Figure 5.16: Frequency Charts of Land Cover Categories	39
Figure 6.1: Probability Surface	45
Figure 6.2: Probability Surface with 50% Threshold	46

List of Tables

Table 2.1: Dependent Variable Limitations in a PMA	10
Table 2.2: Independent Variable Limitations in PMA.....	11
Table 3.1: Project Requirements.....	14
Table 3.2: Phase I Tasks	16
Table 3.3: Phase II Tasks	17
Table 3.4: Phase III Tasks.....	18
Table 5.1: Soil Characteristics	29
Table 5.2: Reclassification of Soil by Drainage and Farmlands.....	35
Table 5.3: Dependent Variable with Topographic Factors	35
Table 5.4: Variance and Eigen Values of the Principal Components.....	36
Table 5.5: LRM Coefficients	40
Table 6.1: Logistic Regression Result	43
Table 6.2: H-L Test of the Final Model	44
Table 6.3: Predicted vs. Observed Values	47

List of Acronyms and Definitions

CAT	Consejo para la Protección del Patrimonio Arqueológico Terrestre de Puerto Rico
CRM	Cultural resource management
DBF	Database file
DEM	Digital elevation model
ED	Environmental determinism
GBD	Geodatabase
GIS	Geographic information system
LRM	Logistic regression model
OGP	Office of Management and Budget of Puerto Rico (OGP)
PCA	Principal Component Analysis
PMA	Predictive model for archaeology
SHPO	State Historic Preservation Office

Chapter 1 – Introduction

This project focused on the creation of predictive models for precolonial archaeological sites in Puerto Rico. Puerto Rico is located in the eastern most part of the Caribbean and is the smallest island among the Greater and Lesser Antilles Islands. It includes various insular territories near the principal islands: Vieques, Culebra, Mona Island, and diverse islets and cays (Figure 1.1).



Figure 1.1: Puerto Rico in Relation to the Caribbean Region

This project aimed to increase the knowledge of the environmental characteristics of archaeological sites in Puerto Rico. In addition, this project sought to find archaeologically sensitive areas—both existing and potential—from a geographic perspective. In Puerto Rico, geographic information systems (GIS) have been applied to study specific archaeological sites, such as at the Caguana site in the municipality of Utuado (Torres & Rodríguez Ramos, 2008). However, the application of GIS on a larger scale analyzing the spatial distribution of archaeological sites in Puerto Rico had not been done.

In Puerto Rico, two government agencies are in charge of protecting the archaeological heritage and endorsing urban projects: the State Historic Preservation Office (SHPO) and Consejo para la Protección del Patrimonio Arqueológico Terrestre de Puerto Rico or Council for the Protection of the Terrestrial Archaeological Heritage of Puerto Rico (CAT). Both agencies have created inventories of archaeological sites and archaeological reports for Puerto Rico. The principal difference between these two agencies is that SHPO is in charge of the administration of federally funded projects and CAT is in charge of implementing national patrimony legislation.

Puerto Rico's geological history has been documented back to the Jurassic Period of the Mesozoic Era (Morelock, Ramírez, & Barreto, 2002). The pre-Columbian history of

Puerto Rico dates back to 4000 BCE. The earliest site documented is Angostura located in the northern central coast area of the main island and known today as the municipality of Barceloneta (Vega, 1999). Many archaeological sites are related to the colonial period, which extends from the Spanish conquest initiated in 1493 to the Spanish American War of 1898 with the arrival of United States troops.

The ecosystems in this region are diverse and play important roles in shaping the spatial patterns of archaeological sites. A predictive model was developed for this project to predict possible archaeological areas in Puerto Rican territory.

1.1 Client

The client for this project was the Acting Director of CAT, Laura Del Olmo Frese. The Instituto de Cultura Puertorriqueña (ICP) that is also named the Institute of Puerto Rican Culture appoints CAT. CAT is in charge of archaeological reports and data. The client provided the geographical data in vector format for the cultural heritage (points) and the archaeological reports (polygons).

CAT has been producing a geographical inventory of archaeological sites and reports since 2008. In 2011, CAT updated this inventory and published a web application for use by archaeologists, historians, and students (Instituto de Cultura Puertorriqueña, 2014).

1.2 Problem Statement

Archaeological sites may not be documented due to problems such as private land ownership, budgetary concerns, and unlawful development practices. When these sites go undocumented, their archaeological values may be lost forever.

In Puerto Rico, there were no attempts to predict the potential archaeological sites using quantitative approaches. This led to the problem of not knowing how to allocate limited resources to search for undocumented sites. Therefore, a predictive model accessible to the client would help identify the possibility of finding archaeological sites that have not been documented officially in certain areas when evaluating new urban development plans.

1.3 Proposed Solution

A predictive model for archaeology (PMA) could provide the essential methodological approach to predicting the areas that are most likely to have archaeological sites. This approach was considered to reduce the necessity of extensive fieldwork that could not be accomplished because of budgetary concerns.

To address the problems discussed in Section 1.2, CAT provided the geographical data required to conduct the desired spatial analyses for this project. CAT also provided additional information, including bibliography information and access to the documentation of archaeological sites.

1.3.1 Goals and Objectives

The principal goal of this project was to design a PMA to predict possible archaeological sites and to evaluate the sensitivity of a place to disturbances from an archaeological perspective.

The objectives of this project were:

- To evaluate the viability of developing a PMA for predicting archaeological sites in Puerto Rico
- To develop a predictive model to analyze the probability of the presence of archaeological sites and patterns of past settlements
- To create a probability surface delimitating archaeological sensitive areas

Reaching these objectives would also promote GIS application in conservation of the archaeological heritage of Puerto Rico.

1.3.2 Scope

The study area is located in the central part of Puerto Rico. It contains major rivers, central mountain chain areas, karst areas, and coastal areas of Puerto Rico (Figure 1.2).

There were 571 sites within the study area. Among those, 274 were recorded as precolonial sites, 165 were recorded as colonial sites, and 132 sites were recorded with no information on periods. The data used for the project were provided in May of 2014. New digitalization from the client after this date was not used in this project.

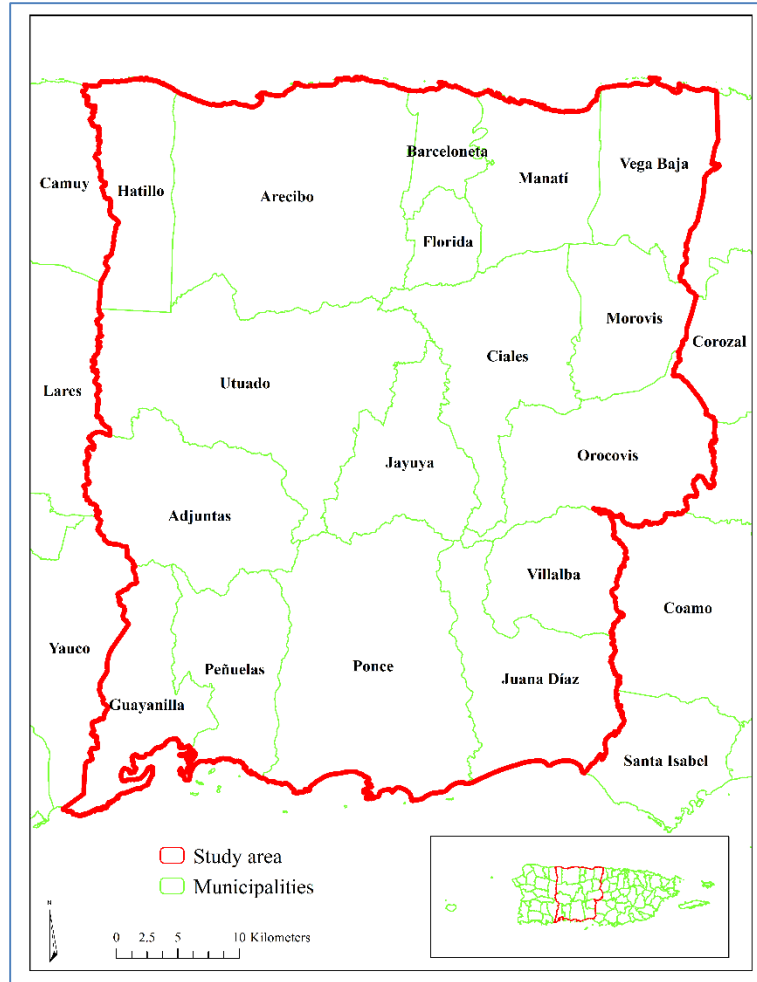


Figure 1.2: Study Area

The precolonial history of Puerto Rico is longer than its colonial history and precolonial sites were analyzed for this project. Analyses considering specific cultural periods or types of archaeological sites were not done due to data limitations. Thus, the precolonial samples were not divided into culture periods such as Pre-Ceramic, Huecoids, or Saladoids. These sub-categories of periods could be considered in the future. Details about the data will be discussed in Chapter 4. The software and tools used were limited to ArcGIS 10.2.2 and Statistical Analysis Software (SAS) for analyzing and creating independent and dependent variables

1.3.3 Methods

This project took the spiral approach to testing the viability of the PMA using different variables. The first phase of the project consisted of scrubbing and creating the independent and dependent variables. This process was accomplished using ArcGIS 10.2.2, ArcGIS ModelBuilder workflows were created to calculate the variables in the study area.

The independent variables included both environmental and social factors. Each factor was represented by a single raster layer in ArcGIS. The DEM was a main data

source to calculate other independent variables such as aspect, slope, and relief. The dependent variable was a point layer with each point representing whether archaeological sites were found in that location.

Once all the data layers were created, both principal component analysis (PCA) and logistic regression analysis were conducted in ArcGIS 10.2.2 and SAS respectively. Using the coefficients calculated from the logistic regression, a probability surface was computed using the Raster Calculator in ArcGIS 10.2.2. Finally, the model was validated using sampled archaeological sites.

1.4 Audience

The principal audience for this project will be archaeologists who use CAT to consult the national archaeological data. This includes students, professors, and the public.

1.5 Overview of the Rest of This Report

The next chapter covers previous work that has been developed using PMA. Chapter 3 discusses how the data were analyzed to create a system to solve the client's problem. Chapter 4 covers the description of the data used to create the independent and dependent variables. Chapter 5 describes the procedures to create the variables and detailed LRM developed. Chapter 6 describes the results and validation processes conducted for the model. Lastly, Chapter 7 covers the conclusions and future work for this project.

Chapter 2 – Background and Literature Review

In Puerto Rican archaeology, most of the applications of geographic information system (GIS) have been for cultural heritage management (CHM) and used to develop maps of archaeological sites. GIS has been used to develop new theories of settlements and the archaeological history of Puerto Rico using GIS (Rodríguez López, 2013). Additionally, GIS has been applied by different agencies in Puerto Rico, including entities such as transportation, environmental studies, and cadaster agencies. The Office of Management and Budget of Puerto Rico (2013) created a web page where it publishes different spatial datasets and web applications. This allows users to access different educational materials and download spatial data of interest.

Although GIS has been applied broadly in other disciplines, its application in archaeology is still limited. In 2013, the Council of Terrestrial Archaeology of Puerto Rico (CAT), in collaboration with the OGP, conducted one of the first national efforts to apply GIS in archaeology. They published a web application in Silverlight to publish information of archaeological sites and cultural resources, accessible to both, the public and archaeologists (Instituto de Cultura Puertorriqueña, 2014). The archaeological data were published after being converted into grids for security purposes, which avoided the disclosure of the location of the sensitive sites. CAT also has hardcopies of archaeological reports and inventories that can be accessed by the public. The web applications allow the examination of archaeological reports and cultural resources from a geographical perspective using maps and aerial images.

2.1 GIS in Archaeology

Archaeologists have used new approaches and new technologies to interpret the history of past societies worldwide. Other disciplines, such as geography and statistics, have also influenced archaeologists to apply different methods to interpret history from a holistic perspective. Geography and GIS have contributed significantly to the discipline of archaeology. GIS has been used in archaeology as a tool to map the archaeological samples and conduct various spatial analyses.

In spite of this, a broad method to interpret the collected samples with GIS is lacking on Puerto Rican archaeology. Wheatley and Gillings (2002) described the term “GISing” (p.1) as the action of collecting big amount of data without developing any further analysis. This is most common in CHM and archaeological heritages offices.

Since the 1970s, cartographic and spatial analysis applications in archaeology increased with the creation of software such as the Synteny Mapping and Analysis Program (SYMAP) (Wheatley & Gillings, 2002). Starting in 1980s, GIS and archaeology began to intersect in their methodologies and approaches with the use of spatial data such as digital elevation models (DEM). This influence of GIS on archaeology was primarily established in the United States, the Netherlands, the United Kingdom, and Europe (Wheatley & Gillings, 2002). Among other applications, integrating GIS into archaeological research started with the predictive models. For example, Kvamme (1988) developed different methods and algorithms to build predictive models. He specified how to calculate the independent and dependent variables. Other authors, such as Hodder and

Orton (1976), described broadly different statistical and spatial applications in GIS, such as point patterns and artifact distributions with regression analysis.

As mentioned before, the predictive model in archaeology (PMA) is one of the first applications of GIS in the archaeology discipline. It was also one of most controversial approaches. According to Wheatley and Gillings (2002), many researchers including Kvamme (1988) had different perspectives about the application of PMA in archaeology. Such scientist considered that predictive modeling is a valuable tool for archaeology.

In the following sections, details of PMA will be discussed, considering theoretical background, statistical approaches, variables, and data used to develop such models. A predictive model can be defined as a mathematical approach to predicting future patterns or facts by analyzing present patterns (Fernández-Cacho, 2009).

2.2 Predictive Models for Archaeology

According to Fernández-Cacho (2009) the difference between the statistical predictive model and a PMA was that the statistical model is used to evaluate the changes on a variable such as a disease, while a PMA uses the variables of existing archaeological sites to predict possible future discoveries. Fernández-Cacho (2009) defined the PMA as: “models to evaluate the grade of archaeological sensitivity areas and evaluates their potential” (p. 9). She further stated that “... predictive model is valuable in archaeology considering that the patterns of settlements of human could be quantified: is possible to evaluate the potential of human settlements for certain territory in the past” (p. 9).

Kvamme (1990) defined a PMA as “an assignment procedure, or rule that correctly indicates an archaeological event outcome at a land parcel location with greater probability than the attributable to chance” (p. 261). Kvamme’s work is one of the most cited in the literature about PMA.

Kvamme (1988) developed methods to calculate variables such as shelter, relief, and aspect. For example, after examining the traditional shelter, Kvamme developed an “interval-level-measure of shelter” (p. 335). This approach and others are discussed further in Chapter 5, with the implementation of some of his methods to calculate variables for the PMA of this project.

2.2.1 Inductive vs. Deductive Methods

One of the main debates regarding applications of PMA is between inductive and deductive methods. The inductive method is commonly described as generalizing information from the specific to the general. When a PMA is based on individual observations to make general relationships, it is described as an inductive method or a data-driven method (Wheatley & Gillings, 2002). According to Fernández-Cacho (2009), the inductive model depends on reliable information to be proven. It was considered mechanic without deep theoretical foundations. Thus, it could be used to do statistical correlations but it lacked interpretation after being conducted.

The deductive method is theory-driven (Wheatley & Gillings, 2002). This method uses known data about the archaeological sites, either empirical or ethnographical (Fernández-Cacho, 2009). The understanding of sites patterns and having a theory a priori, before conducting a PMA is needed. An example, archaeological sites are known to be found in a 500m distance from water resources and in farmland areas. The

deductive method is enriched by statistical analysis to establish the actual presence or non-presence of archaeological sites. With this information, it is possible to identify potential locations of archaeological sites.

Different mathematical approaches are used in both inductive and deductive analyses. The deductive approach often uses functions and map algebra to evaluate certain rules and relationship among variables. In contrast, the inductive analyses depend on the rules that are derived from the characteristics of the data. Methods, such as frequency tables to evaluate the importance of each variable, are applied in inductive studies (Wheatley & Gillings, 2002).

2.2.2 Environmental Determinism

Many authors and archaeologists argued that the PMA is too static. This is because environmental determinism (ED) influences PMA in many ways (Wheatley & Gillings, 2002). ED assumes that human behaviors are influenced by environmental factors, or that environmental factors control human behavior. This school of thought was present in many disciplines including geography and anthropology. A PMA is based on a variety of environmental factors to determine the possible locations of humans, thus, is considered to be influenced by the ED. Scholars such as van Lausen argued that patterns of archaeological sites could be extracted by correlating the presence of archaeological sites with other factors (Wheatley & Gillings, 2002). Other scholars, such as Gaffney (Lock & Stančič, 1995), argued that a PMA is not suitable as it is highly influenced by ED.

Another reason that PMA is considered deterministic is that the independent variables chosen by the researcher are mostly environmental factors. This is because there exist a limitation of cultural variables, which are not commonly produced. On the contrary, the environmental data and archaeological data are produced frequently and accessible for analysis (Wheatley & Gillings, 2002). In addition, the environmental conditions represented by the data in an analysis may not faithfully represent the environmental conditions of past times.

2.2.3 Statistical Regression Methods

Regression methods are often applied in archaeology, including linear regression method and logistic multiple regression (Wheatley & Gillings, 2002). When a logistic regression model (LRM) is applied to conduct a PMA, the independent variables are environmental factors such as elevation, slope, rivers, and creeks (Wheatley & Gillings, 2002). The dependent variable, which is a binary variable, represents the absence or presence of archaeological sites (Wheatley & Gillings, 2002). The presence of archaeological sites is positive data. The absence of archaeological sites is negative data.

However, the negative data could introduce some problems. When a LRM is conducted only with sites (positive data), the negative data are assumed to be simply the absence of positive data. Thus, the negative data are not actual negative data from surveyed areas without archaeological remains (Wheatley & Gillings, 2002). This gap can be filled by having areas surveyed.

2.2.4 Limitations of PMA

Fernández-Cacho (2009) discussed the limitations of independent and dependent variables used in PMA. These limitations are related to the compliance of the data and the archaeological survey approach when the data were collected. These limitations established by Fernández-Cacho (2009) are presented in the following tables, which are based on case studies in Puerto Rico. The limitations established in Table 2.1 are relevant to determining archaeological sites.

Table 2.1: Dependent Variable Limitations in a PMA

Archaeological sites: Dependent variable	
Limitation	Effect
Archaeological site definition	Archaeological site could be defined as a group of remains or isolated remains.
Archaeological tendencies	The methodologies used to collect data in the field by archaeologists could vary. This could affect the chronological periods assigned to the archaeological sites.
Archaeological sites visibility	The changing in topography or anthropogenic impacts on a terrain could impede the documentation of sites.
Methodology	Methodologies applied in the field could differ by the archaeologists
Data entry errors	Errors could occur when entering the location of the archaeological sites or assigning a chronological period

The techniques to record the archaeological information varies. Limitations such as the change government entities, budget problems, and the lack of documentation can affect the collection of archaeological data. For example, regarding data entry errors, different technicians may have different approaches entering information into databases or conducting GIS process. This increases the uncertainty in the documentation of archaeological data.

Throughout Puerto Rican archaeology history, various methods have been used to assign chronological periods to archaeological sites. The most common model to assign relative chronology to precolonial history in Puerto Rico was established by Rouse (1952). Despite the fact that this interpretation model is the most widely used among archaeologists in Puerto Rico, new generations of researchers have established and contributed to new theories of settlements in local and regional chronologies (Rodríguez Ramos, 2010).

Table 2.2 lists the limitations regarding the independent variables (Fernández-Cacho, 2009). These limitations are related to the environmental data and the accuracy of data derived from cartographic materials. According to Fernández-Cacho (2009), most of the

data available are for environmental variables, with very few cultural data. As mentioned before, this is one of the main limitations and most controversial topics about developing a PMA (Wheatley & Gillings, 2002).

Table 2.2: Independent Variable Limitations in PMA

Environmental factors: Independent variable	
Limitation	Effect
Environmental factor are from the present	The environmental factors described the present landscape, not the past landscape
Derived cartography	Variables such as slope or aspect are related to elevation.
Lack of cultural variables	Cultural variables are not frequently produced on geographical data.
Dissimilar information	The information assigned for each archaeological site could be dissimilar. For example, an archaeological site could be situated in a non-agricultural area but very near to other. This spatial information is ignored.

2.2.5 Case Studies

PMA have been applied in many different contexts and at different scales. In the Caribbean, one of the main sources found was the book “Archaeology and Geoinformatics: Case of studies from the Caribbean” by Basil Reid (2008). In this book, various authors discussed diverse cases in the Caribbean and approaches that involved the use of GIS. These studies included cultural resource management and theories about human settlements in the Caribbean.

Reid (2008) discussed the problems in Trinidad about the management of cultural resources given the funding shortage. In order to be able to accomplish studies in archaeologically sensitive areas, a PMA can contribute to the focusing of resources in these areas. He analyzed the presence of archaeological sites using the following variables: watershed, soil texture, land capability, relief, and alluvial plains. The PMA developed by Reid was applied to characterize archaeological sensitivity areas by calculating the probability of finding new archaeological sites. He found that pre-Columbian archaeological sites are most likely to be located in the areas with hilly relief.

The predictive models were also applied in places, including Venezuela (Molina, 2009) and Delaware (Custer et al., 1986). Molina (2009) analyzed a single archaeological site located in Mérida state, evaluating the accessibility for this archaeological site and the optimized routes to access main sources such as water. This analysis was accomplished by using digital elevation models and evaluating characteristics such as environmental zones, height, and economic relations. Molina analyzed the patterns of a single archaeological site to estimate the possible patterns of other archaeological sites.

Custer et al. (1986) developed a predictive model using LANDSAT images. This PMA consisted of environmental zones as independent variables. The LANDSAT satellite images were analyzed to evaluate the different types of environmental zones. First, it was applied the regression method to link the environmental ground areas to the archaeological sites. The study area was then divided into grids to analyze the frequency of archaeological sites by each grid. Probability classes for the archaeological sites were developed and the possible presence of a site in each grid was predicted in this project. The approach of this project emphasized the role of satellite images and the application of statistical logistical regression models.

Another important example of PMA is the Minnesota Archaeological Predictive Model in an area of 12, 872 square kilometers (Minnesota Department of Transportation, 2002). This was one of the most complicated PMA's because it considered different types of variables that normally are not available, such as geological formations. In addition, the model included environmental variables as well as cultural and social variables. This model's results indicated that 86% of observed sites were located high archaeologically sensitive areas.

2.3 Summary

This chapter reviewed the application of PMA, debates, limitations, and its relevance to this project. There are both valid critics and defenders of the PMA. Different methods, such as inductive and deductive, have been applied to develop a PMA. For this project, the main consideration was to develop a PMA able to describe archaeologically sensitive areas.

Chapter 3 – Systems Analysis and Design

The major requirement of the client, the Consejo para la Protección del Patrimonio Arqueológico Terrestre de Puerto Rico (CAT), was to develop a predictive model for archaeology (PMA) to evaluate archaeologically sensitive areas in Puerto Rico. The client is in charge of approving future urban development while protecting archaeological remains from destruction caused by these development activities. Thus, the client sought to streamline the process of evaluating possible impact on archaeological sites in various geographical areas by any future urban projects.

3.1 Problem Statement

Limited funds made difficult conducting extensive archaeological investigation, thus, it is possible that archaeological sites are undocumented in Puerto Rico. On other hand, sites may be documented but for construction purposes, and not as an academic investigation. Low budgetary allocations do not allow CAT to support fieldwork data collection across the entire Puerto Rican region.

Since 2000, in order to protect the archaeological heritage, one of the primary tasks of CAT has been to initiate the digitization of archaeological sites reported to date using geographic information systems (GIS). Continuous updating of new archaeological sites would improve the database for accurate spatial analysis.

3.2 Requirements Analysis

Developing a PMA entailed several functional and non-functional requirements. Table 3.1 lists the requirements that this project addressed. The main functional requirement of the project model was to produce a map of probability surface showing archaeologically sensitive areas in the study area.

Table 3.1: Project Requirements

Type	Requirements
Non-Functional	Data produced by the project should follow archaeological sharing laws in Puerto Rico
	ArcGIS, ERDAS, and SPSS software are required to run the PMA
Functional	The predictive model must evaluate the presence of archaeological sites using variables such as slope, elevation, cost distance to water bodies, soils, forests and others
	The model should produce probability surface ranking locations with the probabilities with the probabilities to find archaeological sites
	Identify significant factors to the PMA
	Archaeological sites should be divided into chronological periods

Other requirements considered in the project were related to the type of data needed to develop the model. The main vector layer was the cultural resources layer representing the archaeological sites or positive sites. This layer was comprised of both precolonial and colonial sites. The other vector data used were the archaeological surveys, which represent positive or negative surveys. This layer was used to extract the negative areas. Additional data, including soils, hydrology, topography, forests, and land use, were required in order to analyze the environmental characteristics of the archaeological sites. Digital elevation model (DEM) data were used to derive additional attributes of archaeological data.

3.3 System Design

Figure 3.1 illustrates the system architecture of this project. The system was based on the data acquisition and its manipulation by different software used to create a final probability surface.

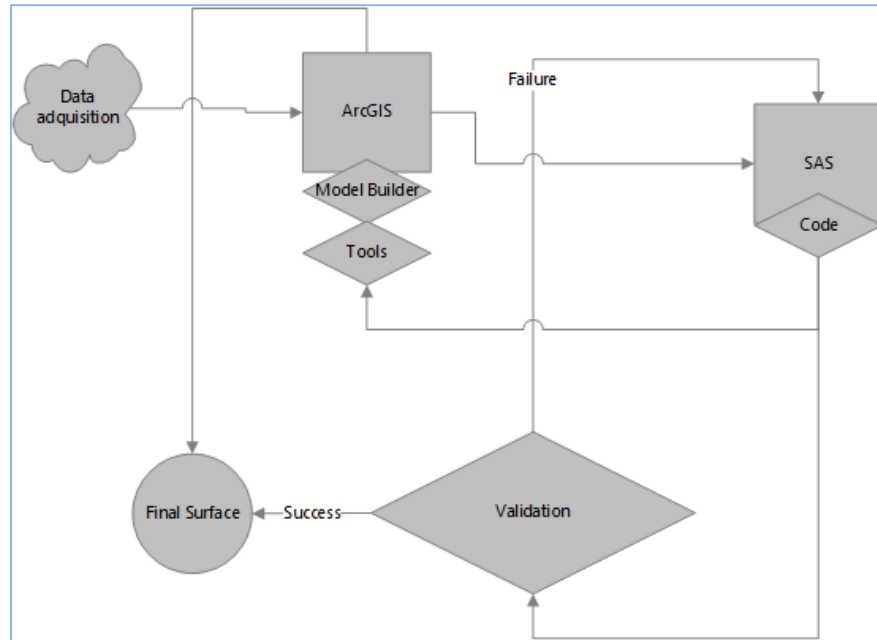


Figure 3.1: System Design

The basic scheme included data acquisition from the client and other agencies. The data were processed in ArcGIS to create the desired variables in a data scrubbing process. After scrubbing process, it was evaluated and calculated independent and dependent variables using ArcMap 10.2.2 software. After having all the data processed and scrubbed, the project advanced to run different statistical analyses. The outputs obtained from the statistical analyses were then used to create the probability surface.

The validation process was conducted using different statistical analyses in Statistical System Analysis (SAS). This was accomplished by using different tools provided by the software and creating models using Model Builder to automate the process of data preparation.

Figure 3.2 shows the data-scrubbing component in the system design. It describes the association of dependent and independent variables after being scrubbed. For any archaeological sites (positive or negative), independent variables, and the dependent variables should be to be linked. A table created with dependent and independent variables were used run the logistic regression analysis and other statistical analysis. After the significant independent variables were identified, a probability surface was created using the coefficients obtained from the logistic regression analysis.

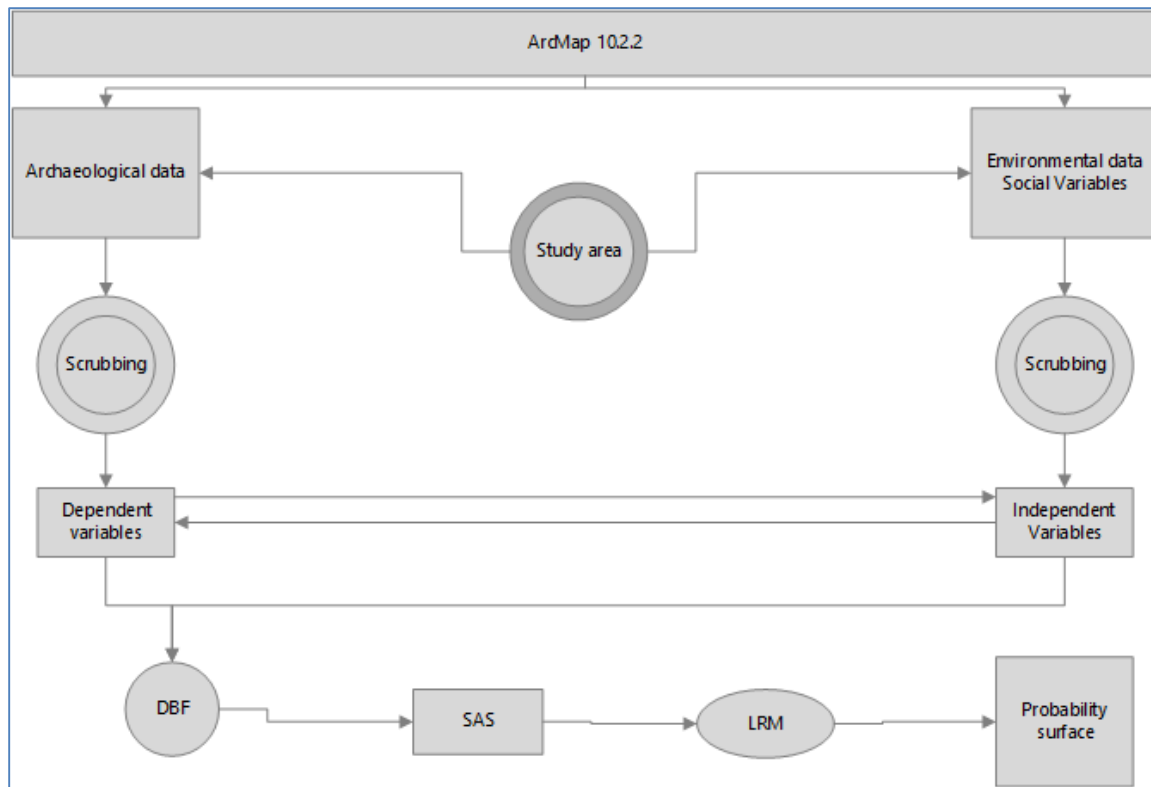


Figure 3.2: Detailed Project Design

3.4 Project Plan

The project was divided into three phases: design, develop and test, and deploy. The development and testing stage was the most intensive, as it entailed the development of all the variables used in the PMA and the statistical analysis.

The first phase — design — consisted of the identification of previous work, external resources for the data, and appropriate models for the PMA (Table 3.2). This was to identify previous case studies of PMA developed in other geographical areas.

Table 3.2: Phase I Tasks

Phases	Task Title
1	Design
1.1	Identifying previous work: literature review
1.2	Identify the data model
1.3	Identify archaeological data
1.4	Identify environmental data

In the second phase, a series of tasks were accomplished (see Table 3.3). The archaeological survey data were cleaned and separated based on chronological periods. In addition, other data representing the independent variables were scrubbed. Different software such as ArcGIS 10.2.2 and SAS were used to create required variables and conduct multiple trials of building the predict model with logistic regression method.

Table 3.3: Phase II Tasks

Phases	Task Title
2	Development and Testing
2.1	Select and collect additional spatial data
2.2	Select geographical study area using main rivers as reference
2.3	Clean the study area data
2.4	Select archaeological sites that are precolonial
2.5	Select negatives archaeological surveys
2.6	Analyze archaeological sites in negative surveys for quality purposes
2.7	Rasterize independent variables
2.8	Create independent variables from DEM
2.9	Create multi-bands raster with the independent variables
2.10	Extract attribute table to table format
2.11	Run the LRM analysis in SAS to assign weights
2.12	Extract logistic coefficients of each independent variable to ArcGIS Raster Calculator tool to create final probability surface
2.13	Map low probability and high probability areas
2.14	Perform quality control analysis for product spatial data

The third and final phase — deploy — consisted of finalizing the model, validating the model results, mapping the results of the PMA, and delivering the project products to the client (Table 3.4).

Table 3.4: Phase III Tasks

Phases	Task Title
3	Deploy
1	Select significant variables to include in the last model using different statistical analysis (LRM)
3.1	Run, improve and test final LRM
3.2	Map probability surface
33	Testing, upgrading and checking the products to deliver
3.4	Organized final geodatabase for the client

3.5 Summary

This chapter described the design of the system to accomplish the goals of the project. After examining the system requirements, the system architecture was developed. Additionally, the project plan was developed to help implement the project.

Chapter 4 – Database Design

This chapter discusses the database development processes involved in the project. Data types used in the project included raster, vector, and tabular datasets. All of these data were stored in an Esri file geodatabase (GDB). The Modelbuilder application within ArcGIS was used to streamline the data processing procedures. The vector data were divided and stored in the GDB as feature datasets. The raster images were stored in the GDB as Raster Catalogs, which organized several raster data into the same catalog. The probabilities and statistical analysis results produced by the logistic regression model (LRM) were stored as tables in the GDB.

4.1 Conceptual Data Model

The conceptual model gives an overview of the entities considered in this project based on a previous model presented by Kadar about data modeling for archaeology (Kadar, n/d). Figure 4.1 illustrates the spatial interactions of these entities.

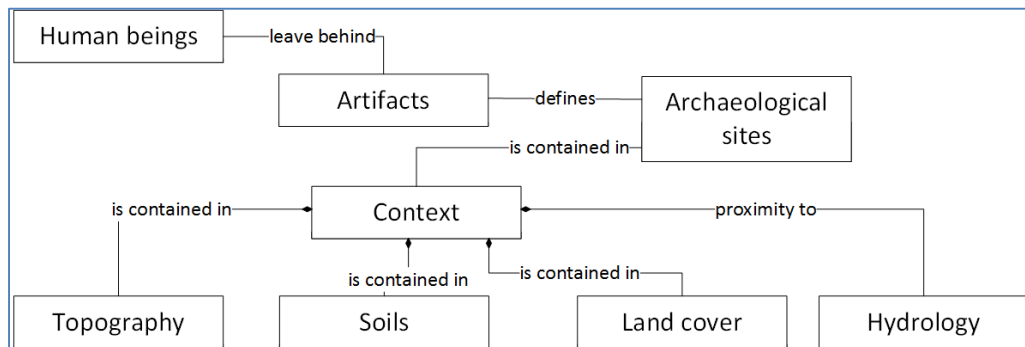


Figure 4.1: Conceptual Model

Human beings left behind materials that in the future becomes evidence for new generations. These materials left behind are called artifacts. A group of artifacts is what makes an archaeological site. Each archaeological site has its own characteristics, defined by the type of remains and the context. The context is defined in this project by the environmental characteristics of a site including topography, land cover, soils, and hydrology. Aspect, elevation, slope, relief, presence of shelter, represented the topography. The soil properties were measured by the drainage and potential for farming. The agricultural areas, urban areas, and forest areas were indicators for land cover. Lastly, hydrology included rivers, creeks, and coastal areas.

4.2 Logical Data Model

The logical data model for this project was based primarily on the types of variables that were chosen for conducting PMA. The GDB for this project contained both vector and raster data (Figure 4.2).

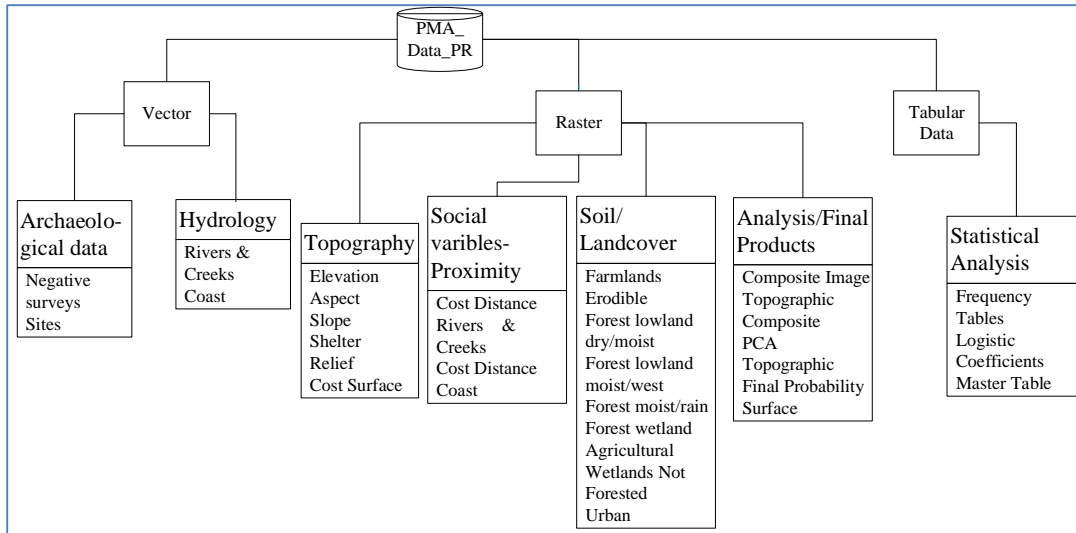


Figure 4.2: Logical Data Model

The vector data included the presence or absence of archaeological sites and locations of rivers and creeks. All of the independent variables were represented in the geodatabase as raster datasets. These included the topographic variables derived from the digital elevation model such as aspect, slope, and relief. Other raster data included the social variables, which are represented by cost distances to different water resources. The cost distance raster was developed to represent how costly it is to traverse an area. The land cover variables of interest to the project were comprised of specific land uses including forests, urban areas, and agricultural land uses. The various outputs were then generated from the different statistical analysis and stored in the geodatabase as raster and tabular data. Figure 4.3 shows an example of how the data was stored. The vector data were organized in feature datasets and raster data were organized using raster catalogs.

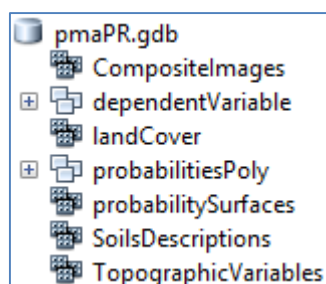


Figure 4.3: Example of Data Stored in the GBD

All raster data were kept individually in the GBD. The tabular data were also inserted in the GBD.

4.3 Data Sources

The majority of the spatial datasets used in the project were downloaded from the Office and Management Budget (OGP) website. A catalog of different types of data was

available on the website: transportation, education, environmental, census, and many others. Due to data restrictions, some datasets, such as orthophotos and topographic maps, could not be downloaded to the local machine. These datasets were accessed using the web server available from the OGP website. The OGP also publishes public data from the federal government agencies, such as Census Bureau and United States Geological Survey (USGS). Most of the data available had metadata. The vector data downloaded from the OGP website were in shapefile format, including hydrology, soils, and legal boundaries in Puerto Rico. The DEM was obtained through a link provided by the OGP.

Land cover data were acquired from the United States Department of Agriculture (USDA) website. These datasets represent the land cover in 1991 and 2000. For the purposes of this project, the 1991 land cover images were chosen. The data downloaded from the USDA websites ((2014) were organized into a single folder (Figure 4.4).

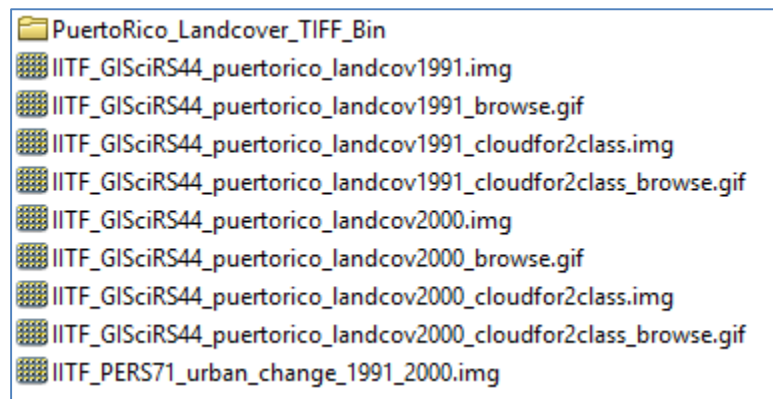


Figure 4.4: USDA Land Cover Data

The data on archaeological sites and surveys, provided by the Consejo para la Protección del Patrimonio Arqueológico Terrestre de Puerto Rico (CAT), had sufficient metadata, including descriptions of the methodologies used to collect and create the data (Figure 4.5).

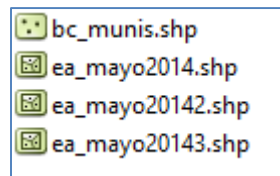


Figure 4.5: Data Received from the Client

The “bc_munis” shapefile records the archaeological sites and “ea_mayo2014” refers to the archaeological surveys. All three “ea_mayo_2014” shapefiles were merged into a single feature class. The main difference between the archaeological sites and the archaeological surveys is that the sites are areas where archaeological evidence has been found and the archaeological surveys are those archaeological studies that have been conducted to ensure the protection of archaeological heritage. The surveys could be either positive or negative, which means finding or not finding archaeological evidence. Despite the fact that many archaeological sites were discovered without the use of surveys, many

others have been documented because of the positive archaeological surveys made compulsory in 1988. (Instituto de Cultura Puertorriqueña, 2014).

4.4 Data Scrubbing and Loading

A data scrubbing process and information extraction for the study area were performed. First, the data were clipped to the extent of the study area. The DEM was smoothed by using the mean elevation within a 100-meter radius. Reducing the noise in the DEM was necessary for deriving other topographic variables such as aspect and slope. From the hydrology dataset, rivers and creeks were extracted. The coastline was digitized using the boundaries of the municipalities.

The archaeological surveys were coded into sites (positive) and no-sites (negative). The sites were coded as 1. The no-sites were extracted from the negative surveys in the “ea_mayo” datasets. The no-sites were coded as 0, representing the areas where no archaeological remains were found. The archaeological surveys, represented as polygons in the original datasets, were converted to points using the Polygons to Point tool.

Further scrubbing of the archaeological data received from the client was conducted. First, the archaeological negative surveys that had sites inside the polygon area were eliminated from the dataset. Negative surveys that were with a 100-meter distance from archaeological positive sites were eliminated as well.

4.5 Summary

This chapter examined the data and the various processes involved in data preparation. Considering the main dependent and independent variables of the project, the file geodatabase structure was chosen. The project used vector, raster, and tabular data. The vector datasets were obtained in form of shapefiles. These datasets were converted into feature classes. Categorized and stored within feature datasets. Raster data received in form of .img format were converted and stored in the file geodatabase as raster datasets in the raster catalog. The final database as described in Figure 4.3 included the data used in analysis and final products of the project.

Chapter 5 – Implementation

This chapter explains the procedures used to develop a predictive model for archaeology (PMA) in Puerto Rico. Calculations of the dependent and independent variables are discussed in this chapter. The data preparation procedures to run the logistic regression model (LRM), such as reclassification of some of the variables, are discussed as well. This chapter also discusses the processes to identify the significant variables in the LRM and the statistical analyses conducted to avoid redundancy. Finally, it is explained how the probability surface was created.

5.1 Dependent Variables

The dependent variable is binary representing whether or not archaeological remains were found at an archaeological site. A value of one indicates positive or the presence of archaeological remains (sites), while a value of zero means the negative or the absence of archaeological remains (no-sites). There are a total number of 274 precolonial sites and 436 no-sites within the study area, which are represented as points in a feature class. Figure 5.1 shows the sites and no-sites within the study area.

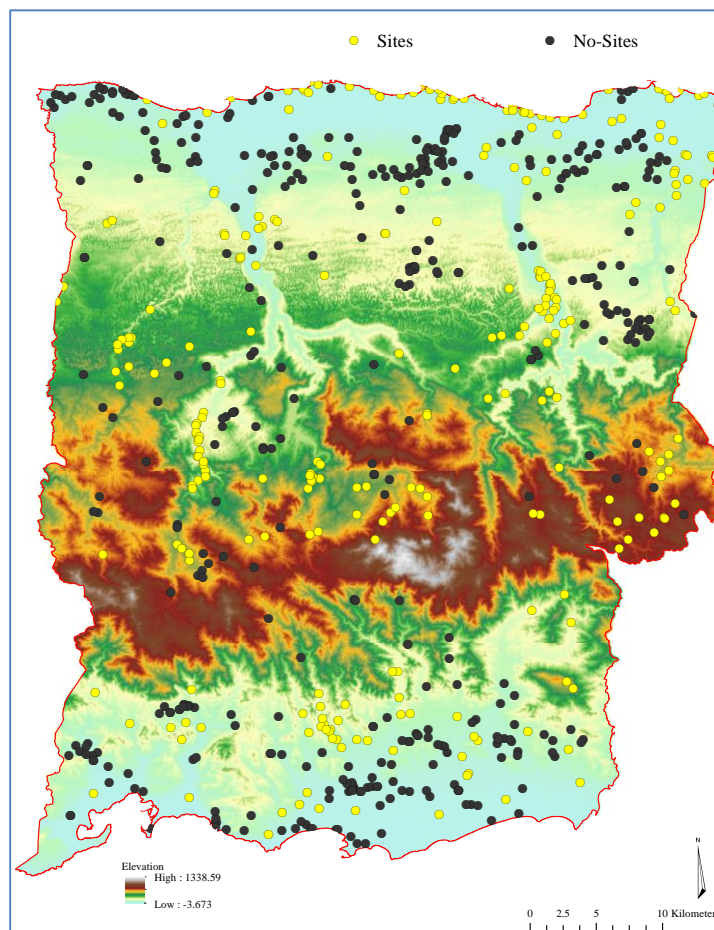


Figure 5.1: Site Samples and No-Site Samples

5.2 Independent Variables

Three types of independent variables were considered: topographic variables, soil and land cover variables, and social variables. Topographic variables such as aspect were used to describe the terrain characteristics that may affect the decision of choosing habitats. For example, it is expected that human beings prefer to choose flat areas that face south. It was also considered the soils and land cover variables, which were characterized their suitability for agriculture and drainage. For social factors, it considered variables such as cost distance to main water resources in the study area.

5.2.1 Topographic Variables

The topography variables (Figure 5.2) were derived from the elevation dataset using the Focal Statistics and Raster calculator tools from ArcGIS 10.2.2. The original DEM provided by the Office of Management and Budget (OGP) is a high-resolution elevation raster with a spatial resolution of five meters.

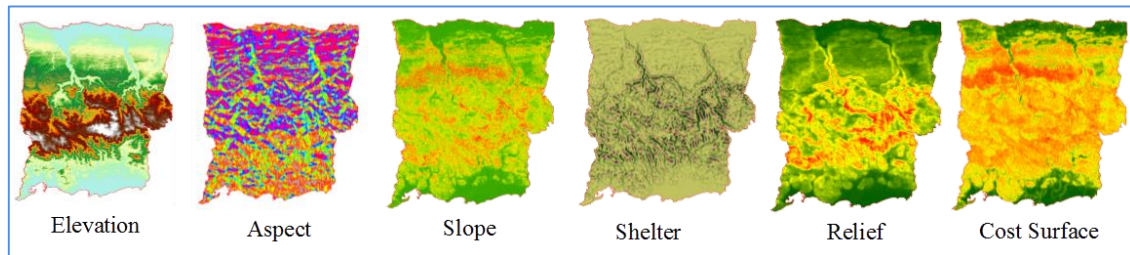


Figure 5.2: Topographic Variables

The elevation was derived from the DEM through a smoothing process. Figure 5.3 illustrates smoothed elevation surfaces. In the study area, mountains are located in the middle area and elevation decreases when moving towards to the coastlines on the north and south.

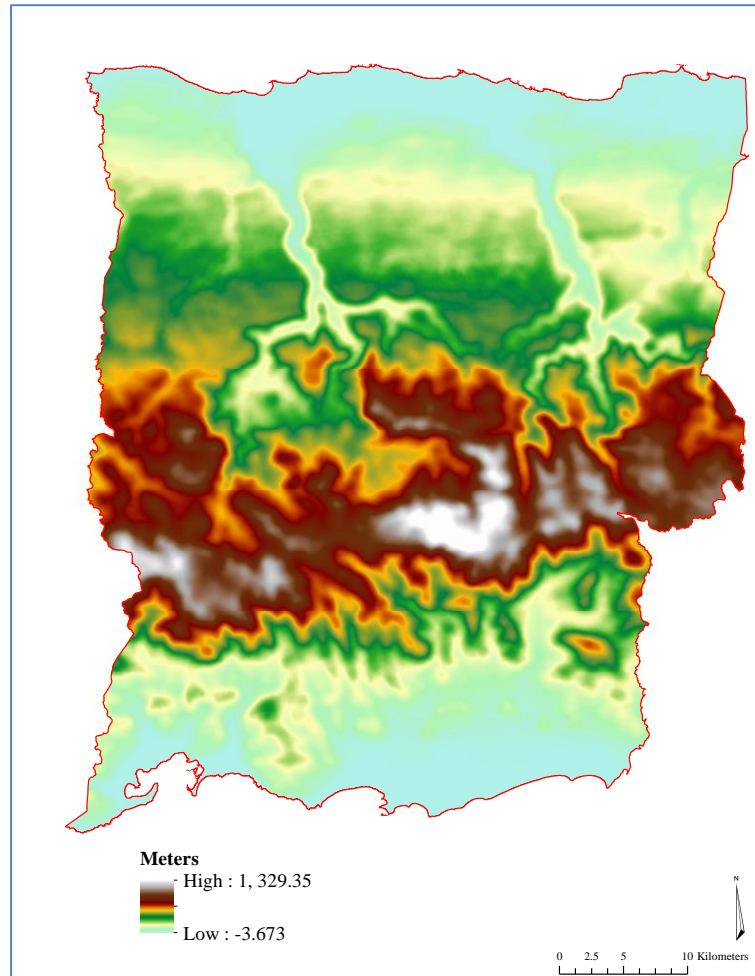


Figure 5.3: Elevation Surface

The steepness of the surface, or slope, is a common factor used to analyze archaeological sites patterns, as found by Kvamme (1988). The slope was calculated using the Slope tool in ArcGIS. This tool identifies the change of elevation over a horizontal distance. The output unit of measurements is in decimal degrees. Figure 5.4 shows that the steepest areas were concentrated in the center of the study area where the highest elevation is concentrated.

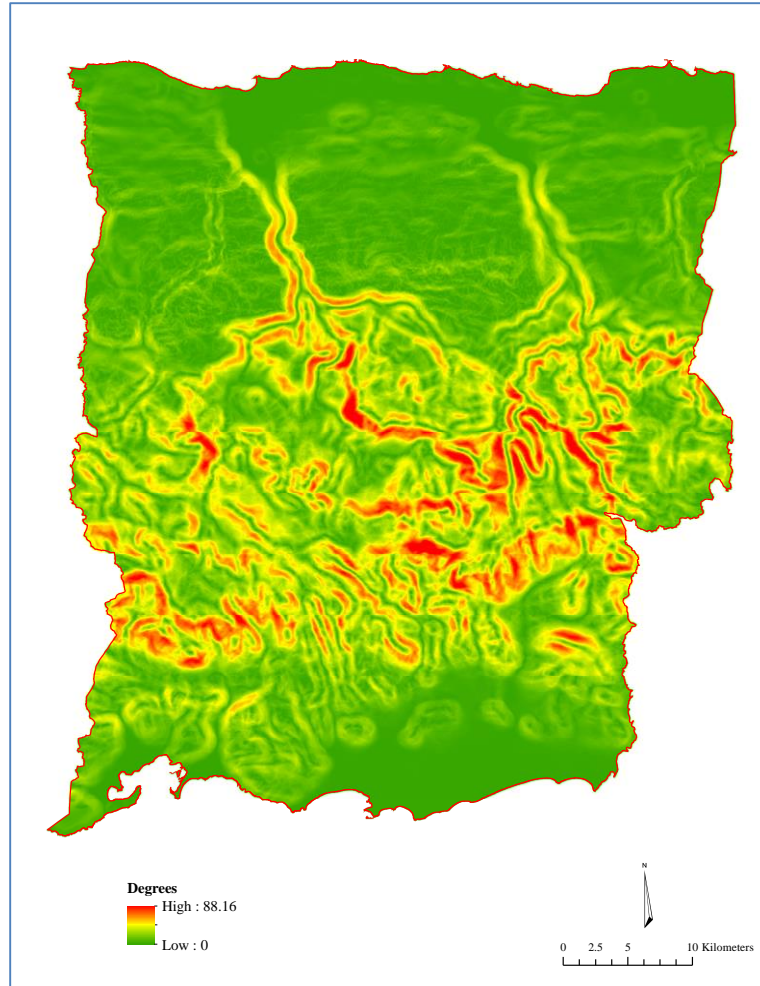


Figure 5.4: Slope Surface

The relief is a measurement of terrain roughness. It describes how abruptly terrain changes across the landscape. This factor is assumed to restrict human activities. The relief was calculated as elevation range of 100-meter radius using Focal Statistics tool in ArcGIS. Figure 5.5 shows the relief output, which is very similar to slope variations.

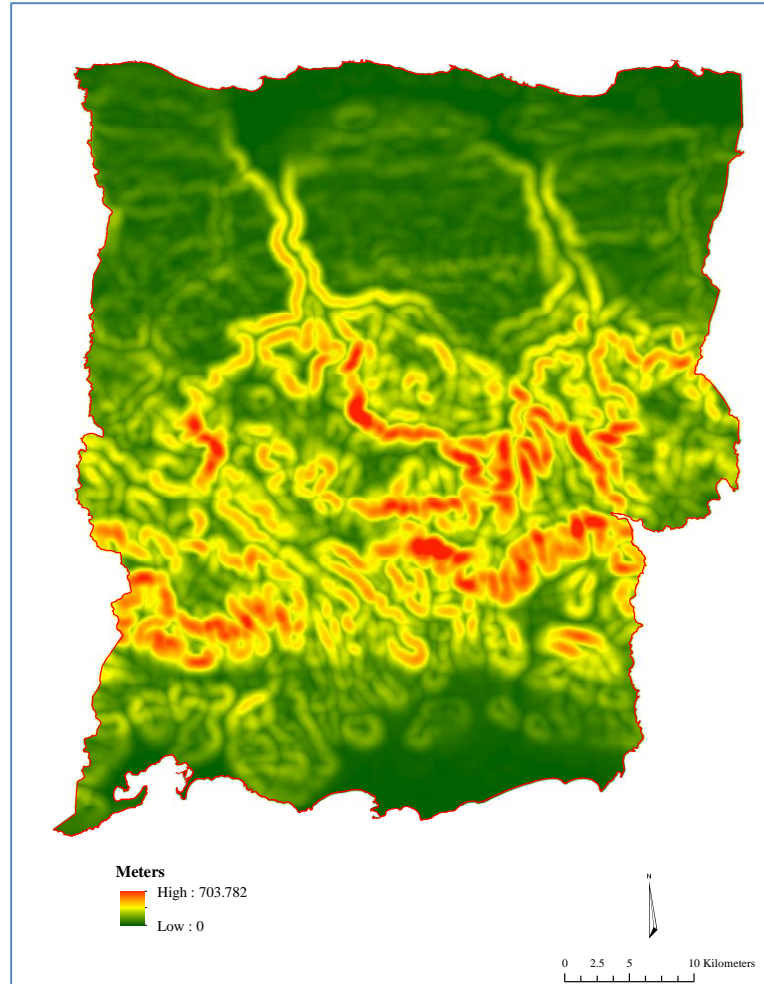


Figure 5.5: Relief Surface

A shelter is defined as an area that can provide refuge from the natural elements (Heilen et al, 2012). The shelter variable was calculated by subtracting the local elevation from the mean elevation. The mean elevation was calculated using the Zonal Statistics tool to assign the mean elevation within 100 meters raster cell. The calculation of the shelter using raster calculator was based on the following formula (Kvamme, 1998):

$$S = \bar{X} - x_i$$

where S represents the difference in elevation between the mean elevation of the neighborhood and the local elevation; \bar{X} represents the mean elevation; and x_i is the local elevation. The negative output values were considered unsuitable locations for shelters while large positive values were considered suitable for shelters (Figure 5.6).



Figure 5.6: Suitability for Shelters

The aspect indicates the exposure to sun illumination. In this project, areas facing the south were more suitable (Kvamme, 1988). First, the aspect was calculated using the Aspect tool. The value in the raster cell values represents the orientation of a location in azimuth, the direction measured to the north.

One hundred eighty degrees were subtracted from the aspect. After obtaining the aspect raster, the aspect in relation with the south was calculated. This process yielded negative and positive values. The sign values were then eliminated by calculating the absolute values. The results were in a range from 0 to 180. The calculation in the Raster Calculator using the following map algebra:

$$SA = |x_i - 180|$$

where SA represents the deviation from the south; x_i represents the local aspect; and 180 represents the south direction. As the original aspect image had values of -1 representing flat areas, those values produced outputs of 181 and were then converted to zero. Figure 5.7 shows the output of the aspect in relationship to the south.

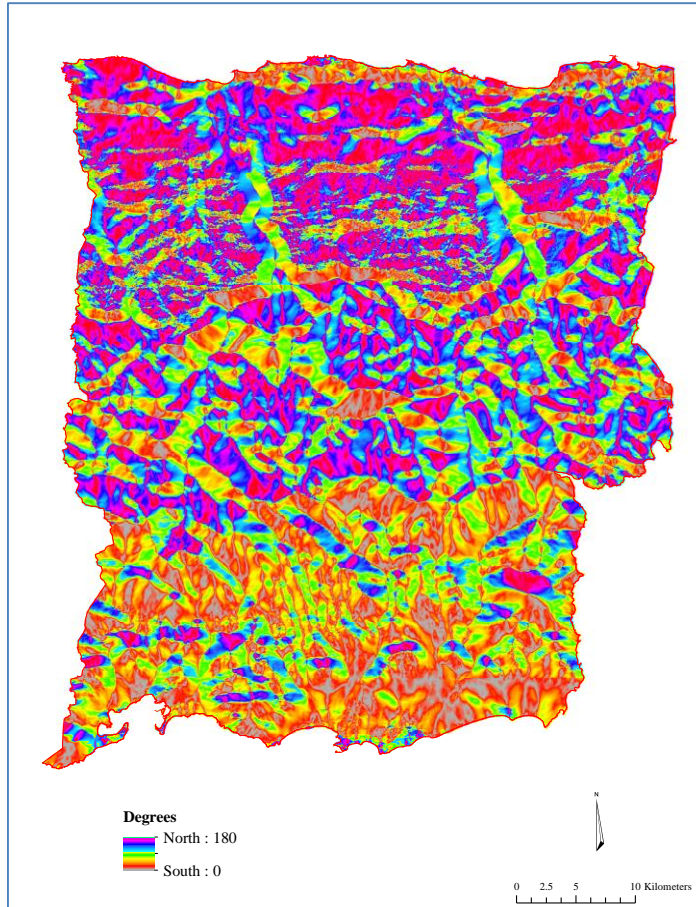


Figure 5.7: Aspect in Relationship to the South

5.2.2 Soils and Land Cover Factors

To obtain land descriptions, two main datasets were used: soils and land cover image. The soils dataset has different descriptions for soil types. The project focused on the farmlands and the drainage descriptions. Table 5.1 describes the fields considered.

Table 5.1: Soil Characteristics

Field	Label	Description
drclassdcd	Drainage Class - Dominant condition	Natural drainage condition of the soil refers to the frequency and duration of wet periods. Dominant drainage.
farmlandcl	Farm Class	Identification of prime farmland, farmland of statewide importance, or farmland of local importance.

Figure 5.8 shows the soil variation by drainage and farmland in the study area. These included the different levels of drainage and different kinds of farmlands assigned by the USDA. The dataset was then converted to raster format using the Polygon to Raster tool in Arc Map. The “null data” in the vector layer was treated as “no data” in the raster.

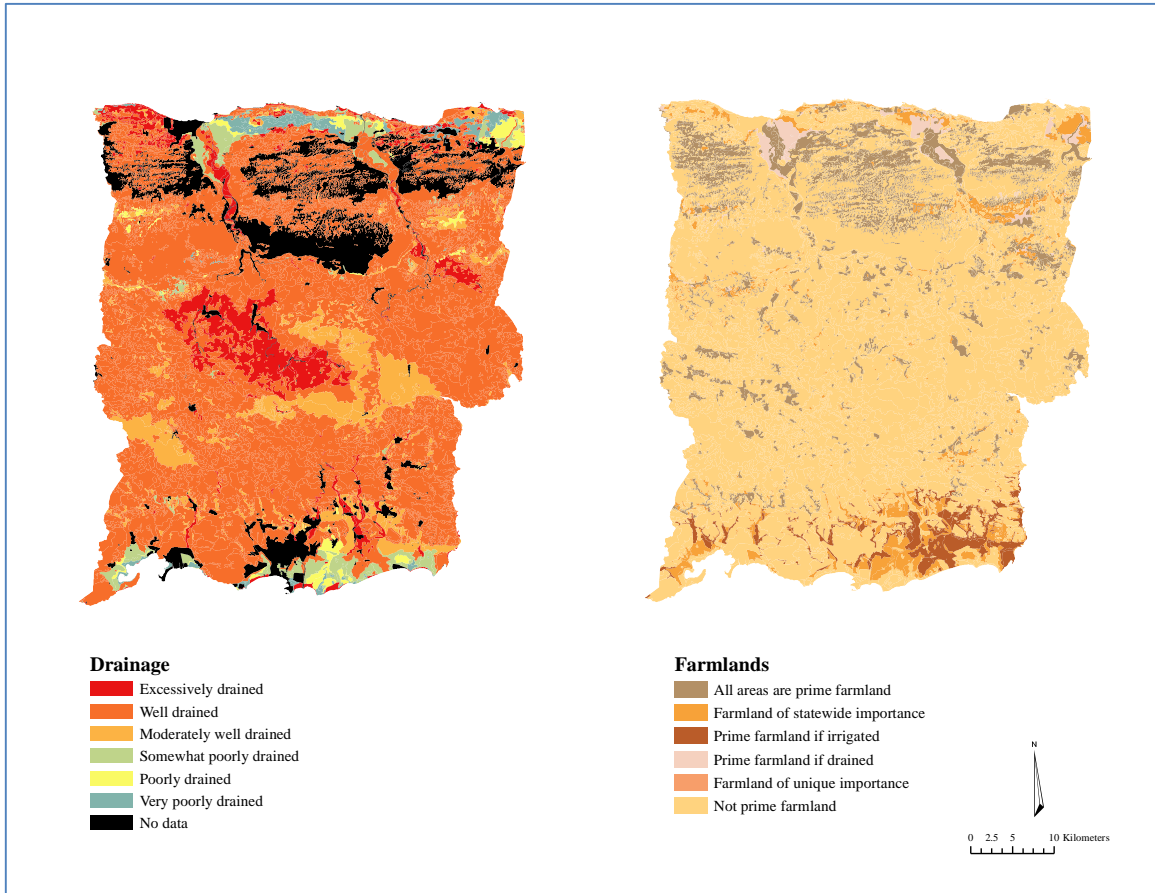


Figure 5.8: Soils Drainage and Farmlands

The land cover data from 1991 (USDA, 2014) were included to study the archaeological sites. The raster data consisted of 28 different land cover types. These twenty-eight types were grouped into seven categories as described in Helmer and Kennaway (2007). The Reclassify tool in ArcGIS was used to classify the original image into seven groups (Figure 5.9).

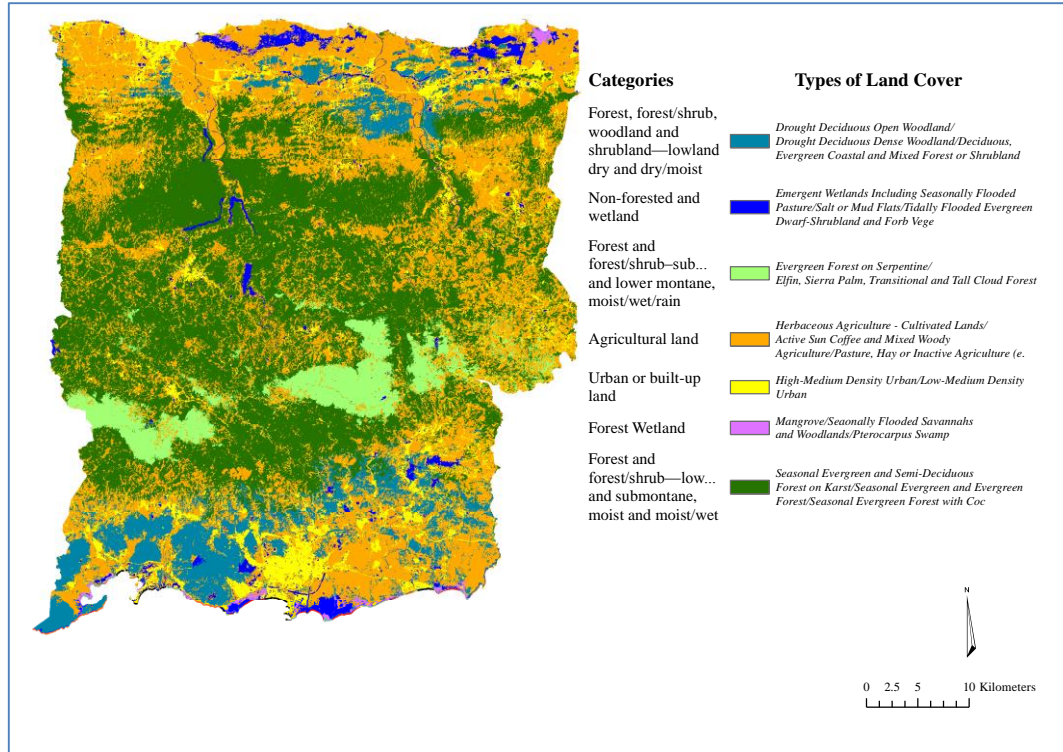


Figure 5.9: Land Cover Data

The categorical data (soils and land cover) in vector format were converted into raster using Polygon to Raster tool.

5.2.3 Social Factors

This project used one social factor: distance to water resources. A cost surface was generated to represent how difficult it was to travel within the project area surface to access water resources. This cost surface was calculated using the natural logarithm of the mean slope in a 100 meter radius and the standard deviation of the elevation in a 100 meter radius (Heilen et al., 2012).

To avoid negative numbers in the result, a value of one was added to each of the variable values before the natural logarithm was calculated as shown below:

$$CS = Ln(\sigma x_i + 1) + Ln(\bar{y}_i + 1)$$

where CS represents the cost; σx_i represents the elevation standard deviation; and \bar{y}_i represents the mean slope. The final product of the cost surface is shown in the Figure 5.10.

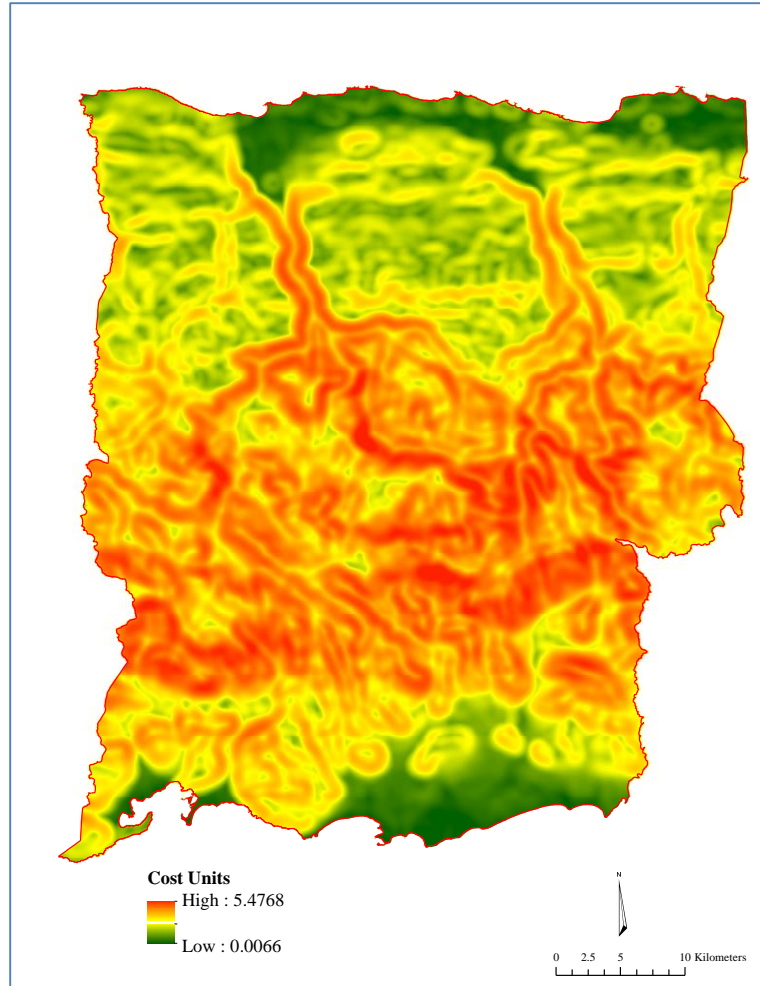


Figure 5.10: Cost Surface

The cost surface was used to calculate cost distance to water resources. The Cost Distance tool “calculates the least accumulative cost distance for each cell to the nearest source over a cost surface” (ArcGIS Resource Center, n/d). The water resources included rivers, creeks, and coast. Figure 5.11 shows the cost distance to rivers and creeks. The lines representing the rivers and creeks showed unexpected breaks in some areas. For purposes of this analysis, only rivers and creeks were extracted from the hydrology dataset. Thus, other types created some gaps on the datasets.

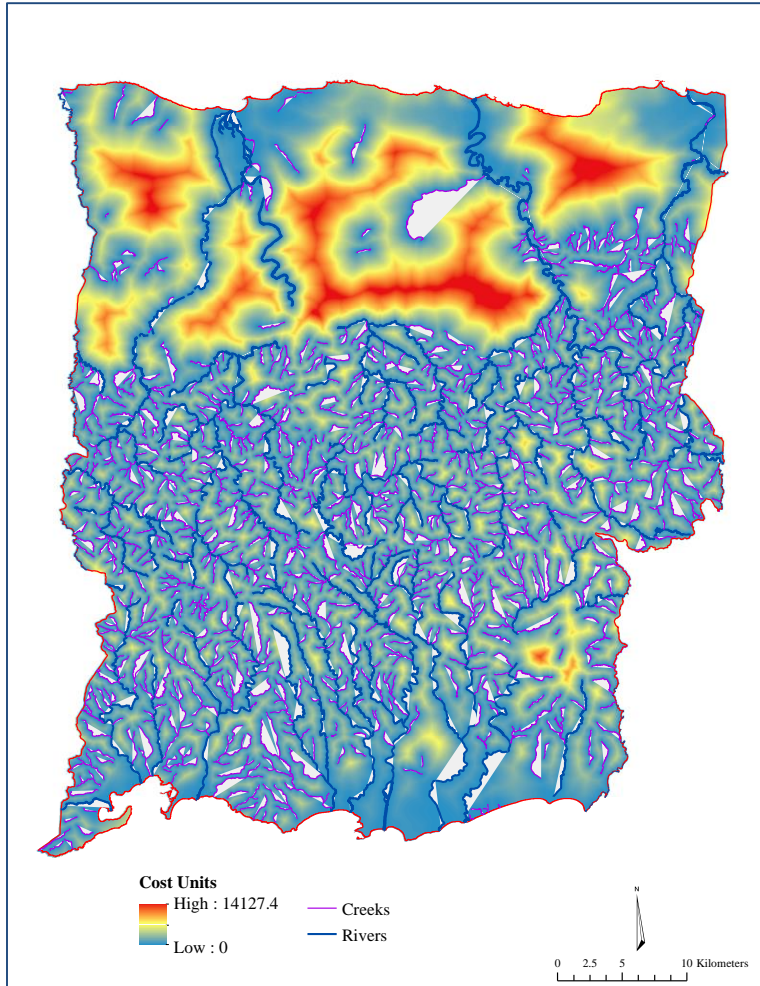


Figure 5.11: Cost Distance to Rivers and Creeks

The same cost surface was also calculated to represent the travel cost to the coastlines (Figure 5.12).

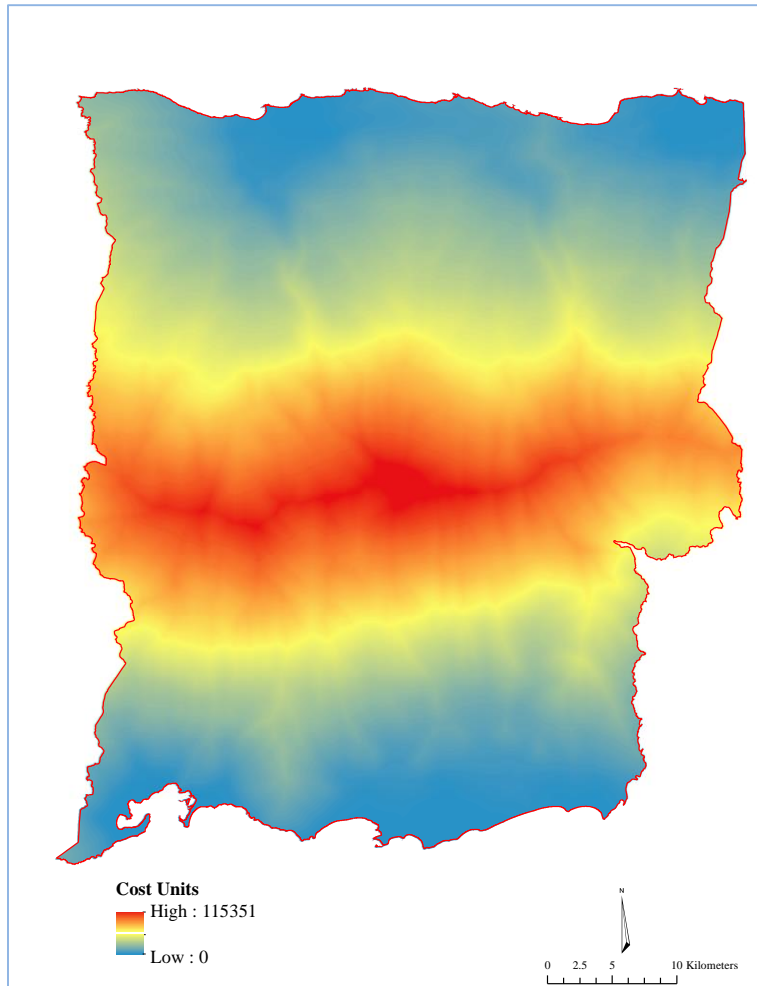


Figure 5.12: Cost Distance to Coastlines

5.3 Re-classifying Categorical Data

The categorical data are nominal measurements such as the soil and land cover types. The categorical variables were then converted into binary values to be used in the LRM. This was accomplished using the Reclassifying tools in ArcGIS 10.2.2. The coding for the soil variables is shown in Table 5.2.

Table 5.2: Reclassification of Soil by Drainage and Farmlands

Drainage	
Category	Binary value
Excessively drained	1
Well drained	
Moderately well drained	
Somewhat poorly drained	
Poorly drained	0
Very poorly drained	
None	
Farmlands	
Category	Binary value
All areas are prime farmland	1
Farmland of statewide importance	
Prime farmland if drained	
Prime farmland if irrigated	
Not prime farmland	0

The land cover variables were also reclassified into binary values. Since there were seven categories of land cover, seven binary raster layers were generated. The categorical variables were converted into raster format using the binary values: 0 or 1. A value of 0 indicates the absence of the variable and 1 indicates the presence of the variable.

5.4 Associating Independent Variables to the Dependent Variable

After creating the independent variables, each of these variables was stacked into a composite raster using the Composite Band tool. The process created an image that included the independent variables as different bands (Heilen et al, 2012). The Extract Multi Value Points tool was then used to assign the values to both sites and no-sites. This process output a table associating each site with independent variables. Sample of the table is shown in Table 5.3.

Table 5.3: Dependent Variable with Topographic Factors

Site (1) No site (0)	Elevation (meters)	Aspect (degrees)	Aspect South (degrees)	Slope (degrees)	Relief (meters)
1	954.38	329.62	149.62	18.25	333.30
0	849.31	88.28	91.71	8.41	99.80
0	838.74	75.90	104.09	6.22	122.01
0	835.50	251.01	71.01	4.90	115.98
1	835.12	337.82	157.82	4.80	82.07
1	833.50	112.33	67.66	4.76	93.98
0	829.19	89.40	90.59	2.44	42.10

The first column represents the dependent variable with the following values: 1 (sites) and 0 (no-sites). This is followed by all the independent variables in each column. The table was exported as a database file (DBF), which was used in the SAS for statistical modeling.

5.5 Identifying Significant Independent Variables

After organizing all the dependent and independent variables in a single table, various statistical analyses were performed prior to building the LRM. These analyses included frequency examination and principal component analysis (PCA). This was followed by a significance test of the variables to evaluate how much a variable affected the likelihood of a location to be a site. After finishing the frequency examination, several LRMs were compared and the one with the best fit was kept.

For the purposes of testing the analysis, a significance level of 5% was used. This was evaluated by running the LRM several times to calculate the significance of the variables. All variables that had significance levels of more than 5% were excluded from the last LRM.

5.5.1 Topographic Variables

It is obvious that topographic variables might be highly correlated among themselves as they were derived from the DEM. To solve this collinearity issue, a PCA was used to reduce redundancy within the independent variables. A PCA creates artificial variables by analyzing the variance and covariance among the independent variables. The PCA was conducted with the seven topographic variables that were stacked in an image. The PCA tool in ArcGIS was used to transform the seven bands into principal components.

The output of the PCA tool also contained seven different components. The component whose eigenvalue was above 1 is normally recognized as important. In this study, eigenvalues were calculated based on the standard deviation (Table 5.4). Based on these results, only the first two components were kept, which contained more than 94% of variance of the original seven factors.

Table 5.4: Variance and Eigen Values of the Principal Components

Principal Component	Standard Deviation	Variance	% Variance	Eigenvalue
1	271.59	73761.12	0.7963	5.5744
2	111.71	12479.12	0.1347	0.9431
3	60.34	3640.91	0.0393	0.2751
4	51.62	2664.62	0.0287	0.2013
5	8.65	74.822	0.0008	0.0056
6	1.58	2.4964	0.00002	0.0001
7	0.72	0.5184	0.000005	0.00003

A scree plot of the eigenvalues is another common tool to determine the important components. Figure 5.13 shows that the first breaking point was at the second component. Thus, all the components after the second one were excluded in the following analyses.

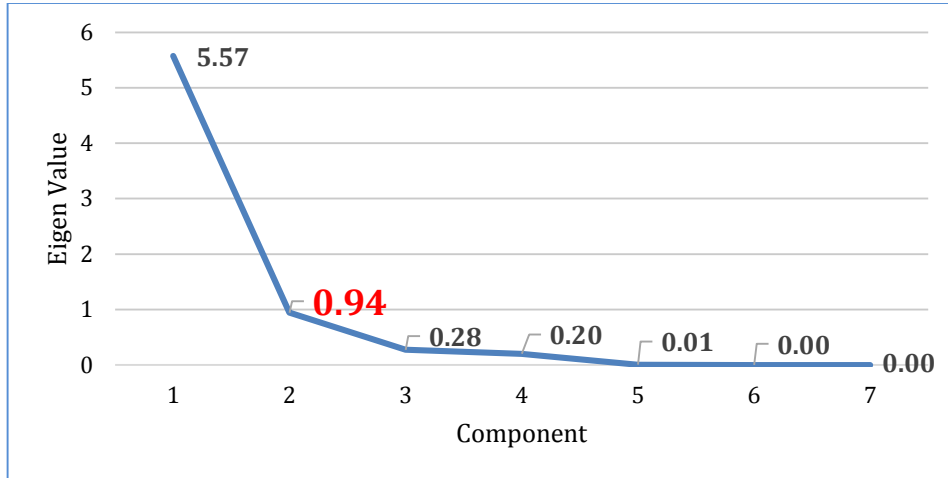


Figure 5.13: Scree Plot PCA Components

A LRM was run for these two variables to evaluate their significance. The LRM showed that only the first component (PC1) was significant with a $p < 0.05$. Therefore, the PC1 was kept in the final LRM.

5.5.2 Soil and Land Cover Variables

The frequency tables were created for the variables of soil and land cover against the dependent variable. The purpose of conducting this analysis was to examine the distribution of land covers and soil types among the sites and no-sites. If the difference between sites and no-sites in a certain category were minimal, the variables would not be significant and thus was excluded. However, when the distribution showed a difference between sites and no-sites, those variables were kept for the following analysis.

The frequency chart in Figure 5.14 illustrates the percentages of sites and no-sites in the farmlands categories. It was observed that 32% of the no-sites fall in this category, as opposed to 11% of sites. This indicates that being primary farmland may significantly influence the likelihood of finding archaeological remains.

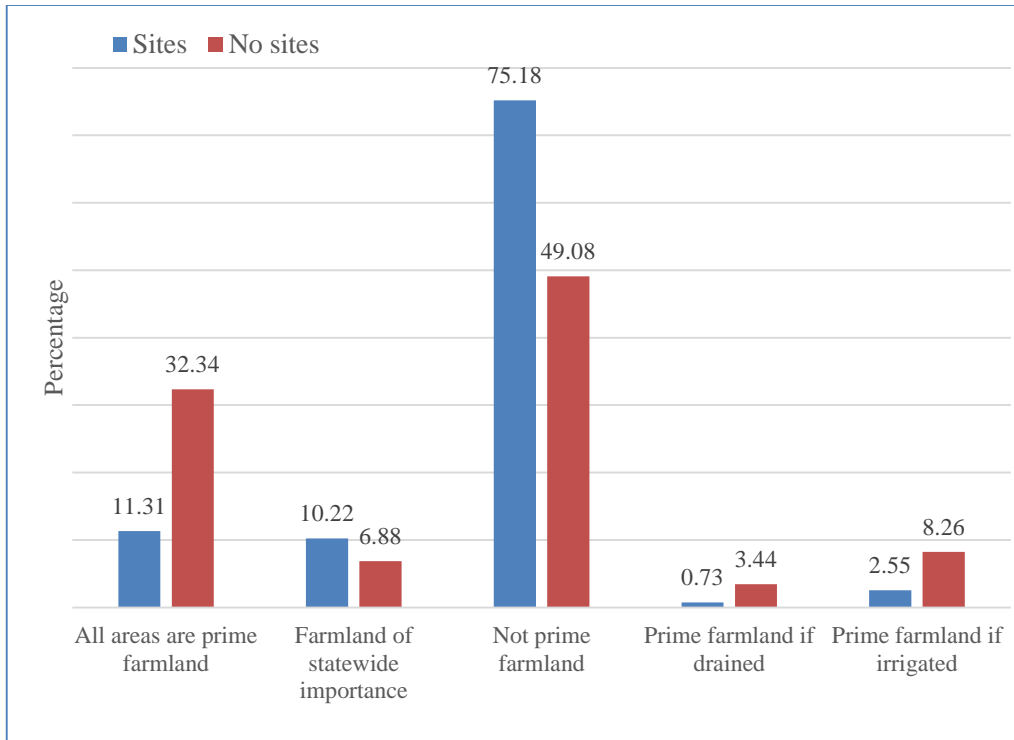


Figure 5.14: Frequency Charts of Farmlands Categories

The same frequency process was followed to analyze the drainage characteristic of the soil. Figure 5.15 illustrates the frequencies of the soil drainage categories. On the excessively well drain areas, 21% of the sites fell in this category while only 10% of the no-sites fell in this category.

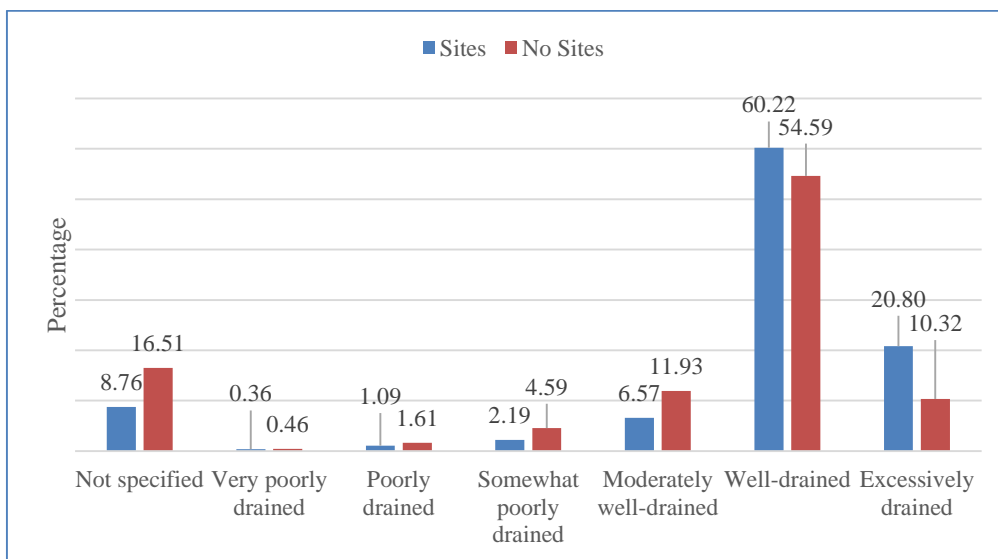


Figure 5.15: Frequency Charts of Drainage Categories

Further, there was a difference of 6% between the site and no-sites in the well-drained category. In total, 87% of the sites were located on areas that are moderately well

drained, well-drained, or excessively well-drained; in contrast, 76% of the no-sites fall in these categories.

A LRM was run separately to evaluate the significance of the farmlands and drainage characteristics. The results showed that the farmlands was significant ($p < 0.001$). Thus, farmland was kept for the final logistic regression model. The same procedure of creating the frequency tables for each of the land cover categories was followed (Figure 5.16).

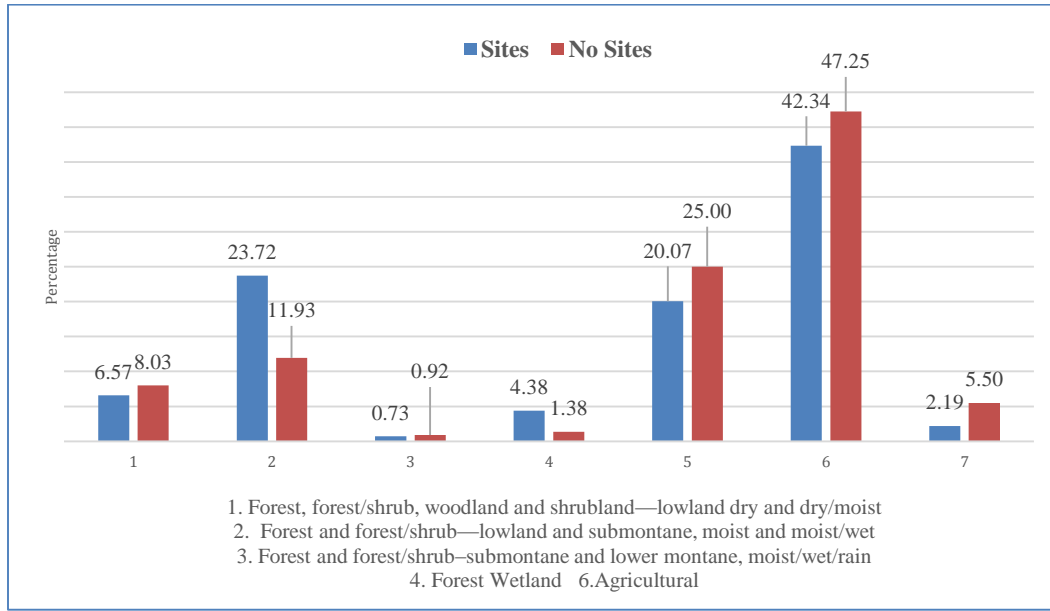


Figure 5.16: Frequency Charts of Land Cover Categories

The category that showed the greatest difference between sites and non-sites was Land Cover 2: “Forest and forest/shrub—lowland and sub montane, moist and moist/wet category.” (Helmer & Kennaway, 2007). This category contained 24% of the sites and 12% of no-sites. To have a better idea of the relationship between the dependent variable and independent variable, a LRM was conducted for the Land Cover variables and only the Land Cover 2 turned out to be significant.

5.5.3 Social Variables

The social variables, cost to various water sources, were also analyzed for their importance using LRM. The results showed that both variables, cost distances to coast and to rivers and creeks, were significant ($p < 0.05$). Therefore, both variables were kept for the final LRM.

5.5.4 Logistic Regression Modeling

The variables that were found to be significant in the preliminary analyses were used to build the LRM using SAS. The included independent variables were cost distance to rivers and creeks, cost distance to coast, farmlands, Land Cover 2 (Forest Moist), and the first principal component (PC1). Table 5.5 shows the regression results where “Estimate”

contains the regression coefficients for each independent variable. Model fitness and coefficients will be discussed in Chapter 6.

Table 5.5: LRM Coefficients

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	1	-0.2829	0.1660	2.9036	0.0884
Cost Distance to Coast	1	0.000010	5.166E-6	3.9052	0.0481
Cost Distance to Rivers and Creeks	1	-0.00011	0.000040	8.2329	0.0041
Farmlands	1	-0.9809	0.1761	31.0168	<.0001
Land cover 2	1	0.7766	0.2296	11.4378	0.0007
PC1	1	-0.00038	0.000753	0.2589	0.6109

5.6 Probability Surface Creation

The probability surface to evaluate the likelihood of presence and no presence of the archaeological sites was created with the logistic coefficients. The probability surface was created using the following equation:

$$p(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \sum_1^n \beta_i x_i)}}$$

where $p(y = 1)$ represents the probability of finding a site (1), e represents the exponential constants (2.71); β represents the logistic coefficients; and x represents the independent variables.

The probability surface was calculated using the Raster Calculator tool in Model Builder. The logistic coefficients were assigned to the corresponding independent variables. The formula used in Raster Calculator was:

$$1 / (1 + \text{Exp} (- 1 * (-0.2829 + -0.00011 * "%Cost Distance to Rivers \& Creeks%" + 0.000010 * "%Cost Distance to Coast%" + -0.9809 * "%Farmlands%" + 0.7766 * "%Land Cover 2%" + - 0.00038 * "%PC1%")))$$

5.7 Summary

Chapter 5 described the process of creating the variables to be used for the LRM. Various group of variables, including topographic, land cover, soils, and social variables, were considered in this project. The categorical variables were prepared and transformed into dummy variables. After all the variables were calculated, each of their values was assigned to the dependent variables: sites and no sites. These values were then transferred to SAS to run the statistical analysis. PCA, frequency analysis, and LRM methods were used to evaluate the significance of the variables. Only the significant variables were used in the final model to generate the probability surface, which was derived using the logistic coefficients of the significant variables in raster calculator.

to SAS to run the statistical analysis. PCA, frequency analysis, and LRM methods were used to evaluate the significance of the variables. Only the significant variables were used in the final model to generate the probability surface, which was derived using the logistic coefficients of the significant variables in raster calculator.

Chapter 6 – Results and Analysis

Chapter 6 focuses on LRM fitness and the final probability surface. In addition, the validation process of the model and its potential to predict the probability of finding a site at a given location is also discussed.

6.1 Logistic Regression Model Fitness

A logistic regression model (LRM) was run with the selected variables discussed in Chapter 5. This significance test showed that all variables were significant except for the first topographic component (PC1) of the topographic variables (Table 6.1). In spite of that, the variable was kept in the model for two reasons. First, topographic variables are recommended in the literature for this type of analysis. Second, the model fitness increases with this variable included, as elaborated below.

Table 6.1: Logistic Regression Result

Parameter	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	-0.2872	0.1660	2.6437	0.0884
Cost Distance to Rivers & Creeks	-0.00011	0.000040	8.2329	0.0041
Cost Distance to Coast	0.000010	5.166E-6	3.9052	0.0481
Farmlands	-0.9809	0.1761	31.0168	<.0001
Land Cover 2	0.7766	0.2296	11.4378	0.0007

Table 6.1 shows the coefficients output of the logistic regression. The logistic coefficients represent the log odds ratios, which represent how much one unit of this variable affects the odds of finding a site. A negative value indicates a negative relationship between the probability of finding a site (1) and the independent variable. On the other hand, a positive coefficient means the independent variable increases the chance of finding a site.

Among the independent variables, it was observed that cost to rivers and creeks, farmlands, and PC1 had negative coefficients. Thus, these variables decreased the probability of finding a site. In contrast, cost to coast had a positive relationship, which meant that it improved the probability of finding a site. All coefficients were tested with the Wald statistics and only PC1 turned out to be insignificant ($p > 0.05$).

The Hosmer-Lemeshow (H-L) test was run to evaluate how well the model fit the data. This test is a goodness of fit test and compares the observed data to the model-predicted values. The null-hypothesis in the H-L test is that there is no difference between observed and model-predicted values. Thus, a significant testing result is desired

as it means that there is no significant difference between the observed and model predicted values.

The H-L of this model had a p-value of 0.27 when PC1 was included, suggesting a good fit of the model to the data; however, the H-L test turned to be significant when PC1 was removed from the model (Table 6.2). In contrast, when the topographic variables were not included, the Chi-Square and p-value increased, which suggested a lack of fitness. Therefore, PC1 was kept in the final model.

Table 6.2: H-L Test of the Final Model

Hosmer and Lemeshow Fit Test			
Final Model with PC1		Final Model without Topographic	
Chi-Square	P	Chi-Square	p
9.9101	0.2714	13.6582	0.0911

A Likelihood Ratio test was also used to validate the model. This test compares the model to a base model that does not include any predictors to evaluate if the model of interest significantly differs from the base model. For this test, a significant result is desired as it shows the base model is significantly improved by including selected predictors. The likelihood ratio test yielded a p-value of 0.0001 with the chi-square value of 92.7 and the degree of freedom of 5. Thus, the model built in this study was valid.

6.2 Probability Surface

Using the coefficients, the final probability surface was created as described in Chapter 5. Figure 6.1 shows the areas more likely to have precolonial archaeological sites, as well as areas with a low probability of such sites.

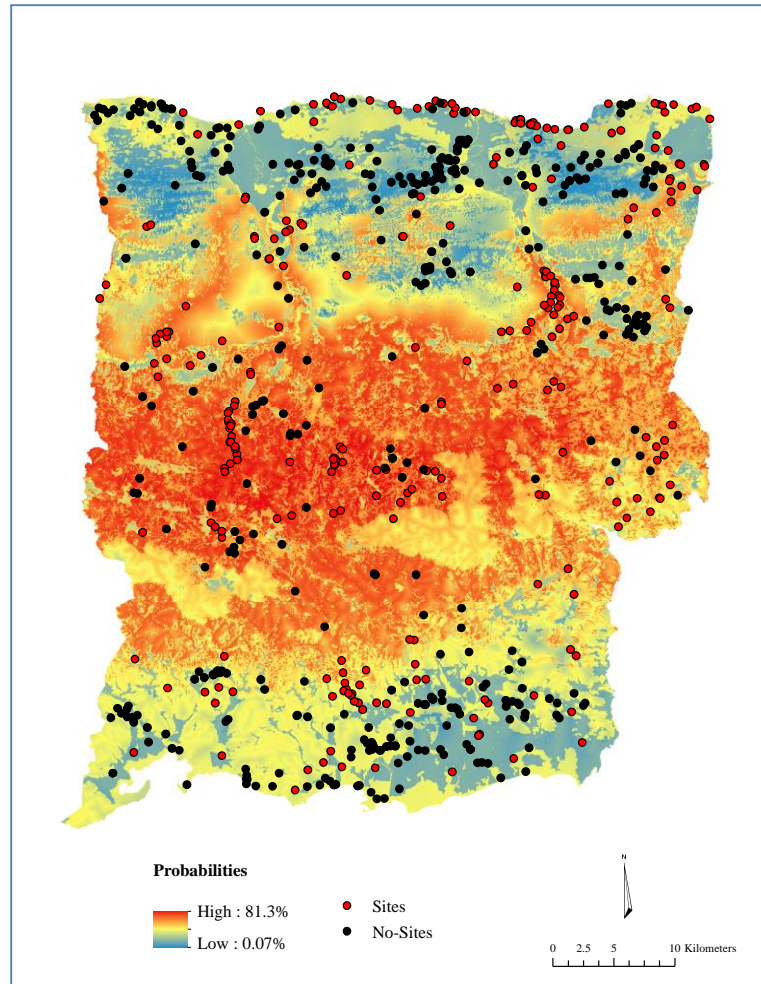


Figure 6.1: Probability Surface

The probability surface suggested probability of finding archaeological remains varied in the study area from 0.07% to 81%. The probability surface was then classified to two classes—areas with high probability and low probability of archaeological sites—by using the most common probability threshold in the literature of 50% (Kvamme,

1990). In the study area, 44% of the total study area had probability values above 50%. In contrast, 56% of the total study area presented probabilities below 50% (Figure 6.2).

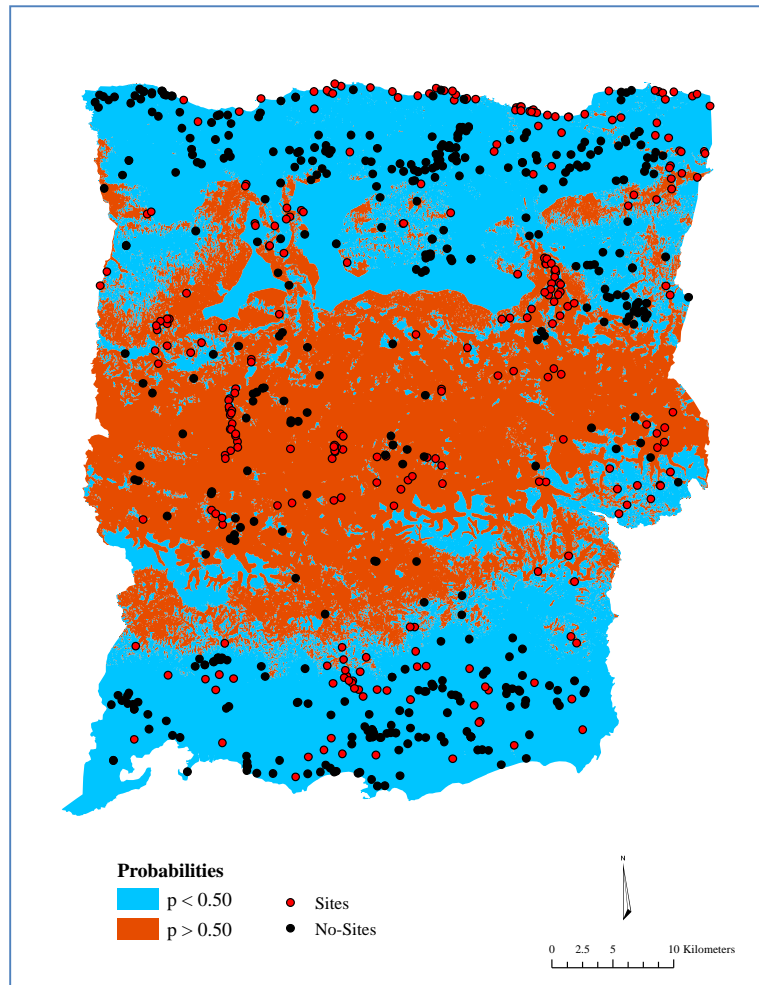


Figure 6.2: Probability Surface with 50% Threshold

Based on the probability surface, the areas that have higher probabilities to find sites are located in the center of the study area. Moving toward to the coast, the likelihood of finding an archaeological site decreases. The probability is used in archaeology as the archaeological sensitivity in the study area (Heilen et al., 2012). Thus, areas with higher probabilities are considered to be more sensitive and vice versa.

6.3 Model Validation

To further validate the model, the observed site and no-site samples were compared to the predicted surface. A threshold of 0.5 was used to divide the probabilities into correctly predicted probability and incorrectly predicted probability. For example, if an archaeological site had a p-value of 0.60, it was considered as correctly predicted. On the other hand, if a no-site had a p-value of 0.60, it was incorrectly predicted. However, if a no-site had a probability of 0.35, it was correctly predicted. For this project, 68% of the

data were correctly predicted according to the classification table of the observed and predicted values (Table 6.3).

Table 6.3: Predicted vs. Observed Values

Probability Threshold	Correctly Predicted		Incorrectly Predicted		Correctly predicted	Percentages
	Sites	No-sites	Sites	No-sites	Total	Correct
0.5	105	378	58	169	483	68

The results show that the model predicts better for the no-sites than the sites. This suggests that when the model predicts a no-site, the probability of being wrong is less. Thus, the model may be better used to locate less sensitive areas for development.

6.4 Summary

The model with the five selected independent variables yielded reasonable fitness and was tested to be valid. With this model, overall 68% of the samples were correctly predicted. However, it seems that model works better for predicting no-sites than sites. This also indicates that other relevant variables might be missing in the model and future analyses will be required to enhance the prediction capability.

Chapter 7 – Conclusions and Future Work

The predictive model for archaeology (PMA) applied in Puerto Rico showed valuable applications and information to evaluate archaeological sensitivity. By examining several variables and their relation to the dependent variable, the project team was able to calculate a probability surface. The model showed that it predicts better no-sites than the sites. This project generated a PMA to help Consejo para la Protección del Patrimonio de Arqueología Terrestre de Puerto Rico (CAT) focus their research and protection of the archaeological remains in a more convenient and efficient way. The PMA proved to be a great tool to evaluate archaeological sensitivity.

The PMA was built using Spatial Analyst and Data Management tools from ArcGIS 10.2.2. The principal component analysis (PCA) was run using ArcGIS as well; the Statistical Analysis System (SAS) was used to obtain the analyses required. Statistical analysis was applied to create the probability surface representing the probability of finding an archaeological site. Several statistical analyses were applied to identify the significance of the variables to be included in the logistic regression model (LRM). The final output was calculated using the Raster Calculator assigning the coefficients from the LRM to each of the selected variables. The product of this process was a probability surface. After creating the probability surface, a Hosmer-Lemeshow (H-L) test was applied to evaluate the predictions of the model. This test showed that the model correctly predicts 68% the total observations within the model.

To finish this project, it is important to establish that it was intended to create a PMA in Puerto Rico for the first time from a regional perspective rather than to a certain sites or smaller areas. Therefore, it is expected to continue improving both the model that was established on this report and the data collected by the government agencies in Puerto Rico respectively. The LRM showed a valuable application to analyze different types of data, either continuous or categorical. Lastly, it was selected to create a LRM because it is one of the most common approaches to developing and conducting PMA because of its flexibility and ease to interpretation.

An area of the model that could be improved is the quality of the data used in the model. Use of high quality and accurate datasets would improve the model and its accuracy. For example, the homogenization of data entry in the archaeological databases will be important for future work. This improvement could allow the analysis of pre-Colonial history of Puerto Rico by its specific periods.

The client's database comprised data containing information about the artifacts found at the sites. Incorporation of this information in a PMA could provide a platform for detailed statistical analysis. As studied by Hodder and Orton (1976), databases of artifacts can reveal patterns inherent within the archaeological history, which in turn can reveal different settlement patterns and processes. This is accomplished by analyzing the frequency of certain artifacts found in sites and analyzing the location of the primary material used to manufacture the artifact, such as type of stone and wood.

The model could also be improved by adding new variables. Other diachronic variables could be added to improve the model, as well as such as temperature and roads. This would provide a platform for analyzing and understanding the impact of urbanization processes and development on the archaeology of Puerto Rico.

Works Cited

- ArcGIS Resource Center. (n/d). Cost Distance. Retrieved 2014, from <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/009z00000018000000.htm>
- Farmer, K. (2008). The Utilization of GIS in the Management of Archaeological Resources in Barbados. In B. A. Reid (Ed.), *Archaeology and Geoinformatics* (pp. 74-86). Alabama: The University of Alabama Press.
- Fernández-Cacho, S. (2009). Bases conceptuales y metodológicas de los modelos predictivos en Arqueología. In *Modelo Andaluz de Predicción Arqueológica* (pp. 8-32). Instituto Andaluz del Patrimonio Histórico: Consejería de Cultura. Retrieved from http://www.iaph.es/export/sites/default/galerias/publicaciones/e-ph-cuadernos/documentos/1254997904589_mapa_silvia.pdf
- Helmer, E. H., & Kennaway, T. (2007). The Forest Types and Ages Cleared for Land Development in Puerto Rico. *GIScience & Remote Sensing*, 44(4), 356-382.
- Hodder, I., & Orton, C. (1976). *Spatial Analysis in Archaeology*. (D. L. Clarke, Ed.) London: Cambridge University Press.
- Instituto de Cultura Puertorriqueña. (2014). *Consejo para la Protección del Patrimonio Arqueológico Terrestre de Puerto Rico*. Retrieved from <http://www.icp.gobierno.pr/programas/consejo-de-arqueologia-terrestre>
- Kadar, M. (n/d). Data Modeling and Relational Database Design in Archaeology. *Acta Universitatis Apulensis*. Retrieved from http://www.uab.ro/auajournal/upload/52_461_modelareadb.pdf
- Kvamme, K. (1988). Development and Testing of Quantitative Models. In J. W. Judge, & L. Sebastian (Eds.), *Quantifying the Present and Predicting the past: Theory, Method, and Application of Archaeological Predictive Modeling*. Denver, Colorado: U.S. Department of the Interior Bureau of Land Management.
- Kvamme, K. (1990). The fundamental principles and practice of predictive modelling. In A. Voopris (Ed.), *Mathematics and information science in archaeology: a flexible framework* (pp. 112-125). Bonn: Holos-Verlag: Studies in Modern Archaeology.
- Lock, G., & Stančić, Z. (1995). *Archeology and Geographical Information Systems*. London: Taylor & Francis Group.
- Minnesota Department of Transportation. (2002). *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota*. Minnesota.
- Molina, M. O. (2009, May-August). Use of Geographical Information Systems (GIS) for the development of predictive models in prospective archeological survey. *Red de Revistas Científicas de América Latina, el Caribe, España y Portugal*, 27(76), 219-236. Retrieved from <http://www.redalyc.org/articulo.oa?id=71218616004>
- Morelock, K., Ramírez, W., & Barreto, M. (2002). The World's Coasts: Puerto Rico. *Coastal Papers*. Retrieved from <http://geology.uprm.edu/Morelock/wcpr.htm>
- Oficina de Manejo y Presupuesto de Puerto Rico. (2013). *Portal Datos Geográficos Gubernamentales*. Retrieved from <http://www.gis.pr.gov/>
- Reid, B. A. (2008). *Archaeology and Geoinformatics: Case Studies from the Caribbean*. (B. A. Reid, Ed.) Tuscaloosa, Alabama: The University of Alabama Press.
- Rodríguez López, J. A. (2013). *Apuntes para la historia de la arqueología en Puerto Rico*. San Juan, Puerto Rico: Ediciones Puerto.

- Rodríguez Ramos, R. (2010). *Rethinking Puerto Rican Precolonial History*. Tuscaloosa, Alabama: The University of Alabama Press.
- Torres, J. M., & Rodríguez Ramos, R. (2008). The Caribbean: A Continent Divided by Water. In B. A. Reid, *Archeology and Geoinformatics: Case Studies from the Caribbean* (pp. 13-29). Tuscaloosa, Alabama: The University of Alabama Press.
- United States Department of Agriculture. (2014). *Natural Resources Conservation Services: Caribbean Area*. Retrieved from Prime Farmlands Definitions: http://www.nrcs.usda.gov/wps/portal/nrcs/detail/pr/soils/?cid=nrcs141p2_037285
- Vega, J. (1999). *Resultado de laboratorio de radiocarbono: Sitio Angostura, Barceloneta, Puerto Rico*. Submitted to Consejo de Arqueología de Puerto Rico.
- Wheatley, D., & Gillings, M. (2002). *Spatial Technology and Archaeology: The Archeological Application of GIS*. New York: Taylor & Frances Group.

Appendix A. Sample Code SAS

This is a sample of the code used in SAS to run the statistical analysis.

```
1 data main;
2 inputs OBJECTID decision Elevation Aspect AspectSout Slope Shelter Relief CostSurfac
Cost Distance to rivers and creeks Cost Distance to coast
3 cards; /*variables values excluded*/
4 run;
56
/*Significance for categorical variables*/
7 proc logistic data=main (DROP=Elevation Aspect AspectSout Slope Shelter Relief Cost
Distance to rivers and creeks Cost Distance to coast CostDistMa
8 class Farmlands Land Cover 2/ param=ref;
9 model decision=Farmlands Land Cover 2 / rsq lackfit ctable pprob = .5;
10 run;
11
12 /*Significance for topographic variables*/
13 proc logistic data=main (DROP= Cost Distance to rivers and creeks PC1 Cost Distance to
coast CostDistMa ForestRain Farmlands Erodible ForestDry
14 model decision=Elevation Aspect AspectSout Slope Shelter Relief CostSurfac /
selection=stepwise rsq lackfit ctable
15 run;
16
17 /*Significance for social variables*/
18 proc logistic data=main (DROP= Elevation Aspect AspectSout Slope Shelter Relief
CostSurfac ForestRain Farmlands
19 model decision=Cost Distance to coast Cost Distance to rivers and creeks /
selection=stepwise rsq lackfit ctable pprob
20 run;
21
22 /*Last model with selected variables: PC1*/
23 proc logistic data=main (DROP=CostDistanRiversOnly Elevation Aspect AspectSout Slope
Shelter Relief CostDistMa CostSurfac
24 class Farmlands Land Cover 2;
25 model decision=Cost Distance to coast Cost Distance to rivers and creeks Farmlands
Land Cover 2 PC1 / rsq lackfit ctable pprob = .5;
26 run;
27
28 /*Last model with Elevation*/
29 proc logistic data=main (DROP=CostDistanRiversOnly PC1 Aspect AspectSout Slope Shelter
Relief CostDistMa
30 model decision=Cost Distance to coast Cost Distance to rivers and creeks Farmlands
Land Cover 2 Elevation / rsq lackfit ctable pprob = .5;
31 run;
32
33 proc print data=main;
34 run;
```