PROBABILITY MODEL FOR ARCHAEOLOGICAL SITE LOCATION, A CASE STUDY ON O‘AHU ISLAND, HAWAI‘I

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

GEOGRAPHY AND ENVIRONMENT

APRIL 2019

By

Laura L. Gilda

Thesis Committee:

Qi Chen, Chairperson
James M. Bayman
Yi Qiang

Keywords: likelihood, GIS, environmental attributes, cultural resource management

ABSTRACT

The objective of this research was to create and implement a locational probability model to identify activity areas utilized by past populations that could still contain remnant or intact cultural resources in unsurveyed areas. Consideration of human behavior, observable patterns of land use, and observable landscape modifications can contribute to locations of culturally utilized locations and areas. A wider perspective could provide a more reliable landscape analysis and interpretation of a wider range of features. A successful model would benefit decision making for land use management, development, cultural resources management (CRM), and preservation. With today's environmental funds reducing and the increase of construction and expansion, identification of archaeological and cultural resources is necessary in the remaining unsurveyed areas to establish baseline information required for informed management decisions and long-term stewardship (Zeidler 1995).

The goal was to determine the most useful environmental factors for identifying probability areas that accurately predict utilized or underutilized areas across the landscape and incorporate them into a probability model as applied for land and resource management using Geographic Information Systems (GIS), StatView, R, and MATLAB. For this study, previously identified cultural resource data on O‘ahu USAG-HI controlled lands served as a case sample.

TABLE OF CONTENTS

ABSTRACT ii

LIST OF TABLES vi

LIST OF FIGURES vii

LIST OF ABBREVIATIONS viii

CHAPTER 1: INTRODUCTION 1

STATEMENT OF PROBLEM 3

RESEARCH GOALS 3

RESEARCH STRUCTURE 5

CHAPTER 2: STUDY AREA 6

ENVIRONMENTAL OVERVIEW 6

ENVIRONMENTAL VARIABLES 8

CHAPTER 3: BACKGROUND AND PREVIOUS STUDIES 9

DEVELOPMENT OF SITE LOCATION MODELING 9

LITERATURE REVIEW 10

Archaeological Site Location Prediction and Probability Modeling Overview 10

Archaeological Site Location Prediction in Hawai'i 12

Lessons Learned by Past Studies 14

ARCHAEOLOGICAL BACKGROUND 15

Settlement and Agricultural Intensification Patterns on O'ahu 16

CULTURAL RESOURCES MANAGEMENT ON USAG-HI LANDS 18

CHAPTER 4: PROJECT METHODOLOGY 20

CULTURAL RESOURCES DATA: TRAINING AND TEST SAMPLES 20

Training Sample 21

Test Sample 21

CREATION OF RANDOM SAMPLE 21

SELECTION OF ENVIRONMENTAL VARIABLES 23

Aspect 23

Elevation *24*
Rainfall *24*
Slope *25*
Distance to a Stream: Fresh Water *25*
Solar Radiation *26*
Soil Type *26*
Land Cover *27*
DATA PROCESSING 27
STATISTICAL ANALYSIS 29
Descriptive Statistics *29*
Inferential Statistics *31*
MODELING 44
Logistic Regression *44*
Performance Assessment *45*
Model Output *48*
MODEL VALIDATION 51
CHAPTER 5: CONCLUSION 55
CHAPTER 6: DISCUSSION AND FUTURE DIRECTIONS 59
DATASET LIMITATIONS 59
ADDITIONAL ENVIRONMENTAL ATTRIBUTES 60
UTILIZATION OF OTHER IMAGERY 61
ADDITIONAL MODEL VALIDATION 61
FURTHER DIRECTIONS 61
REFERENCES CITED 62

LIST OF TABLES

Table 1. Summary of Variables…………………………………………………………….28

Table 2. Descriptive Statistics for Environmental Variables for Sites (N=730)………………...29

Table 3. Paired Means Comparison H_o=0 of Numerical Variables of Sites Verses Non-Sites….32

Table 4. Chi-Squared Tests for Categorical Variable……………………………………………32

Table 5. Coefficients Output for Environmental Variables in R for N=1460…………………...45

Table 6. Table of Deviance………………………………………………………………………46

Table 7. Summary of K-Fold Cross Validation Iterations………………………………………47

Table 8. Descriptive Statistics for Validation Sample (N=89).………………………………….51

LIST OF FIGURES

Figure 1. Cultural Resource Survey Areas on Installations with Training Sample Sites…...……7

Figure 2. Cultural Resource Survey Areas on Installations with Random / Non-Site Points…....22

Figure 3. Continuous Environmental Attribute Histograms (a-e) with Normal Curve Distribution……………………………………………………………………………30

Figure 4. Categorical Environmental Attribute Histograms (a-c)………..……………………..31

Figure 5. Aspect Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………………34

Figure 6. Elevation Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………………35

Figure 7. Rainfall Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………………36

Figure 8. Slope Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………………38

Figure 9. Distance to Stream Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………39

Figure 10. Solar Radiation Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………40

Figure 11. Soil Character Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………41

Figure 12. Land Cover Class Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites………………………………………………………………………43

Figure 13. Plot of the ROC Curve and AUC of 0.76…………………………………………48

Figure 14. Predicted Likelihood Map for O'ahu Island………………………………………50

Figure 15. Distribution Comparison of Test Validation Points, Sites (a, N=89) and Non-Sites (b. N=80)……………………………………………………………………………..52

Figure 16. Survey Results Test Sample Sites and Non-Sites Overlaid with Probability Zones....53

Figure 17. USAG-HI Installations Showing Unsurveyed Areas Within the .68 – 1.0 Probability Zone………………………………………………………………………………57

LIST OF ABBREVIATIONS

Acronym	Definition	Acronym	Definition
AEC	Army Environmental Command	MMR	Makua Military Reservation
ANOVA	analysis of variation	MRLA	Major Land Resource Areas
ArcSDM	species distribution model	MUKEY	map unit key
AUC	Area under the curve	NAD83	North American Datum 1983
ccap	Coastal Change Analysis Program (also C-CAP)	NED	National Elevation Datum
CHILD	Channel-Hillslope Integrated Landscape Development	NOAA	National Oceanic and Atmospheric Administration
CRA	Common Resource Area	NOS	National Ocean Service
CRM	Cultural Resources Management	NRCS	National Soil survey Center
DEM	Digital Surface Model	PARC	Pililaau Army Recreation Center
DSM	Digital Elevation Model	ROC	receiving operating characteristic
DTM	Digital Elevation Model	SAR	Synthetic Aperture Radar
ERS	European Space Agency	SAS	Analytics Software and Solutions
ESRI	Environmental Systems Research Institute	SBMR	Schofield Barracks Military Reservation
GDB	geodatabase	SDSFIE	Spatial Data Standards for Facilities, Infrastructure, and Environment
GIA	graphical intuitive approach	SOEST	School of Ocean and Earth Technology
GIS	Geographic Information Systems	SRTM	Shuttle Radar Topographic Mission
GPS	Global Positioning System	SSURGO	Soil Survey Geographic Database
HARN	High Accuracy Reference Network	TMS	Thematic Mapper Simulator
HoLIS	City and County of Honolulu, Honolulu Land Information System	USAG-HI	United States Army Garrison, Hawaii
ICRMP	Integrated Cultural Resources Management Plan	USDA	United States Department of Agriculture
IfSAR	interferometric synthetic aperture radar	UTM	Universal Transverse Mercator
KTA	Kahuku Training Area	UXO	Unexploded Ordnance
LiDAR	Light Detection and Ranging	WGS84	World Geodetic System 1984
lulc	land use/ land cover	WH/m2	watt hours/ meters squared
MATLAB	Matrix Laboratory		

CHAPTER 1: INTRODUCTION

Modification of the physical landscape results in a cultural landscape, according to Carl Sauer (1925). These modifications from human land use over time result in archaeological or cultural resource locations. Ancient and modern population centers have consistently occupied coastal regions and expanded inland. Identification of these areas of cultural activity in densely vegetated and variable terrain remains a challenge to cultural resource managers around the world. General settlement patterns seem to transcend time. What is not understood, is if these locations can determine if particular environmental attributes such as terrain, soil composition, and access to water have contributed to people selecting locations to conduct their activities.

This study examines eight environmental variables using spatial analysis of geographic locations. The goal is to determine if locations of known past human activity can be identified and differentiated in a systematic way to anticipate where other activity areas may be. The case study location is the island of O‘ahu, utilizing United States Army Garrison, Hawaii (USAG-HI) cultural resources data from 15 different installations of the 35 on O‘ahu.

How people have interacted with the changing physical environment contributes to the evolving cultural landscape, as theorized by Carl Sauer. Better understanding of how humans interact with the physical environment and its conditions and each other through evolving economic and social development could provide broad-scale insight on land use patterns and systems. Combinations or layering of permanent environmental variables are likely to identify places of past human activity. Spatial analysis can determine how preference of attributes influence human behavior and impact the natural environment resulting in observable patterns.

Analysis of environmental setting and spatial patterning in areas that have been previously researched can provide insight for understanding areas less studied. Being able to anticipate what activities took place and where on the landscape they took place is important for land use management, development planning, and archaeological research. Locations where evidence remains of human activities are known as cultural resources and archaeological sites, referred to as “sites” hereafter. Material or structural remnants provide information on what types of activities took place at these sites, the resources utilized and sometimes the time period and duration of use, as well as the setting can provide information on environmental qualities that made that location desirable.

Current cultural resource management (CRM) practices rely on accurate identification of sites across the landscape. Site discovery procedures in rural and difficult to access regions rely on surveys of selected land parcels conducted on foot (pedestrian surveys) which are time consuming, labor intensive, expensive, often hazardous, and limited by intensity of survey coverage and spatial perspective. Development of a successful archaeological site location probability model would be extremely useful for CRM to prevent or mitigate damage or destruction from proposed land use practices and development. It would also lend understanding of how past site location selection was influenced by various interconnected environmental factors from a landscape perspective (Connolly and Lake 2006).

Hawaiian archaeology is largely focused on identification of surface architecture, the residual stone structures of past lifeways. A wider perspective could provide a more reliable landscape analysis and interpretation of a wider range of features. Consideration of human behavior, observable patterns of land use, and observable landscape modifications can contribute to locations of culturally utilized locations and areas. A successful model would benefit decision making for land use management, development, CRM, and preservation. In addition to understanding how environmental variables are related to human settlement patterns, it is also important to accurately identify archaeological resources to serve the broader interests of cultural heritage and academic communities. With today's environmental funds reducing and the increase of construction and expansion, identification of archaeological and cultural resources is necessary in the remaining unsurveyed areas to establish baseline information required for informed management decisions and long-term stewardship (Zeidler 1995).

Technological advancements and software enhancements that can be performed with geographic information systems (GIS) and remote sensing data such as aerial photography and Light Detection and Ranging (LiDAR) are opening avenues to new ways to identify sites across large tracts of land, areas of dense vegetation, and remote locations faster and more cost effectively. These improvements have expanded the ability to define parameters for geological, ecological and topographical attributes found consistently and with recognized patterns. This research will explore development of activity area location probability or suitability modeling to provide ways to reduce survey time and energies, streamline ground truthing efforts by identifying likely areas for sites that are targeted for development, and indicate areas to avoid until a pedestrian survey can be undertaken. The research objective is to develop and implement a locational probability model that predicts areas likely utilized by past populations that could still contain intact or remnant cultural resources in unsurveyed areas.

Statement of Problem

A local model for Hawai´i has not been created. Most early spatially analytical models utilized three geophysical and environmental variables consisting of elevation, slope, and aspect (Dalla Bona 2000). A Cultural Resource Management Plan prepared in 1998 (Anderson) for USAG-HI, reviews basic logic and common knowledge to project where sites may be found. These predictions focused on narrow bands of land along streams and level plateaus. These physiographic landforms such as gulches, plateaus and steep slopes were used to estimate high, moderate and low probability. Later Integrated Cultural Resources Management Plans (ICRMP) funded by the Army were never finalized and never expanded much beyond Anderson's basics (Ziedler 2009).

This research aims to identify common environmental attributes that coincide with known site locations and use this knowledge to develop a model that can predict where site locations may be in unsurveyed areas.

Research Goals

The goal of the current study will be to determine the most useful environmental factors and socioeconomic needs for identifying probability areas that accurately predict utilized or underutilized areas across the landscape and incorporate them into a probability model as applied within GIS. For this study, previously identified cultural resource data on O‘ahu USAG-HI controlled lands will serve as a case sample upon which to base the predictive model.

Questions to be addressed include:

1. Can any environmental variables be identified as more important to locations of human activity?

2. How can GIS and remote sensing techniques help to identify potential activity areas and contribute to understanding human interaction with the environment in the past?
3. Can a useful predictive probability model be created for locating cultural resources or their sub-types?

This study focuses on the identification of high probability areas for the presence of past human activity or sites. These types of sites include Traditional Hawaiian sites such as agriculture and habitation sites, Historic era sites such as ranching, plantation agriculture, and military activities. Some of these site types overlap locations or may have transitioned continuously over time.

While a model will not eliminate the need for pedestrian surveys, it is hoped that reasonably accurate predictions can be beneficial for management decision making. Confidence in high probability areas may influence planning decisions to either move project areas or realize that pedestrian surveys need to be conducted before a project gets too far along. Sites will still have to accessed to be properly recorded and evaluated.

For CRM on USAG-HI Installations, successful probability prediction greater than 70% would be desirable. Successful identification of sites 70% of the time is used as a baseline confidence measure. Confident estimation that sites are likely to occur in the identified area is sufficient for land use and development planning and justifying funding for surveys to confirm site presence or absence. The type, or intensity of survey, may be tailored to the predictive likelihood once accuracy of the model is confirmed. Once sites are identified, they can be treated as appropriate and the land use or development efforts can be adjusted or advanced. Early anticipation and identification efforts at the beginning of development planning streamlines project planning and facilitates cost savings.

In addition to elevation, slope, and aspect identified in previous studies, exploration of several other geophysical and environmental features will be evaluated. Factors such as proximity to fresh water sources, solar radiance, soil type, and precipitation will be considered for independency, correlations, and causal relationships between each other as well as with sites. Non-site attributes will also be considered to determine if characteristics can be narrowed down.

Research Structure

For this study, a set of environmental attributes was compiled for sites on the O‘ahu USAG-HI lands, which resulted from archaeological identification surveys and construction monitoring projects. Observations of elevation, slope, rainfall, topography, soil types, hydrology, and slope aspect were examined for 730 site locations. These attributes are largely unchanged over time, apart from rainfall which has changed somewhat, however the waterways remain largely unchanged. A combination of these observations was used to identify patterns towards creation of a probability model for sites locations. The sample site data does not appear completely random; there is a variety of conditions that appear consistent while others do not, such as fairly level slopes. These data were largely collected with submeter accurate global positioning systems (GPS) and ground-truthed by staff in the field. Data analysis for the known O‘ahu USAG-HI dataset for USAG-HI controlled lands provides useable variables for similar upland unsurveyed areas used for training area and infrastructure development planning.

Pedestrian surveys in Hawai‘i are largely focused on observations of constructed surface structures or land modifications remaining from past human activity. These surveys do not identify subsurface or buried sites however, the environmental attributes may indicate surface characteristics and the potential for buried deposits may be assessed through archaeological testing. Analysis of these attributes may contribute to the modeling structure. Aside from identifying parameters that contribute to the model, there are attributes that will be unique and not contribute to definition of a pattern or be considered as a model constraint. It is anticipated that attributes such as degree of slope, proximity to water and other resources and soil type could contribute strongly to modeling. Consideration of these attributes in relationship to the known functional use of that site could contribute to future modeling and an enhanced or new understanding of land use over time.

The presentation of this study is as follows: (1) A review of the background and previous studies of location modeling for archaeological sites; (2) The study area and environmental variables will be discussed as well as the methods for selecting environmental variables; (3) The data processing and statistical methodology will be reviewed and analysis presented; (4) The results and conclusions will summarize the success of the project and provide some insight on challenges encountered and suggestions for future work on the topic.

CHAPTER 2: STUDY AREA

Environmental Overview

The Hawaiian Islands are located roughly in the center of the Pacific Ocean. The island chain is isolated, with the main eight islands located roughly 2,480 miles from the North American west coast, 3,100 miles south of Alaska, 4,100 miles east of Japan, 3,850 miles northeast of the Mariana Islands, and 2,600 miles north of the Polynesian Islands and Marshal Islands. O‘ahu is the third largest island in the Hawaiian archipelago. O‘ahu formed from three shield volcanoes, Wai‘anae, Ka‘ena, and Ko‘olau (Sinton et al 2014), which have since eroded into two steep sided parallel mountain ranges with a broad saddle area between them. Later eruptions formed several tuff and ash cones on the southern coastal plain. Being approximately 3 million years old, perennial and intermittent streams have cut the mountains and plains into numerous valleys and plateaus. The island perimeter is fringed by reef on the eastern and southern sides and otherwise is surrounded by deep waters. These attributes created a very fertile uplands and lowlands and rich and diverse marine surroundings.

The study area selected is limited to lands managed and controlled by the USAG-HI. Just over 1000 cultural resources between the sea level at Pililaau and Fort DeRussy (1 meter (m)) to the summit of Mount Ka´ala (1202.4m) in Schofield Barracks have been identified on 21 installations since the beginning of the Army’s presence on O‘ahu in 1909. These sites include traditional Hawaiian, historic ranching, sugar plantation, and early military. Cultural resources data for 730 sites with trusted location information were utilized from fifteen installations on O‘ahu with site information already in GIS format. Pedestrian surveys to identify cultural resources have been conducted over 23.8%, or approximately 14,217.9 acres, of the total 59,652.8 acres (USAG-HI 2018) of USAG-HI managed lands (Figure 1). These data are in a GIS database maintained and managed by USAG-HI.

Many USAG-HI lands have been historically modified through not only ranching and pineapple cultivation, but also military operations. As a result, few native species of vegetation persist in the lowlands or in areas dominated by introduced grasses, drought-resistant trees and Polynesian introductions. Today the uplands are largely undeveloped lands but are dominated by

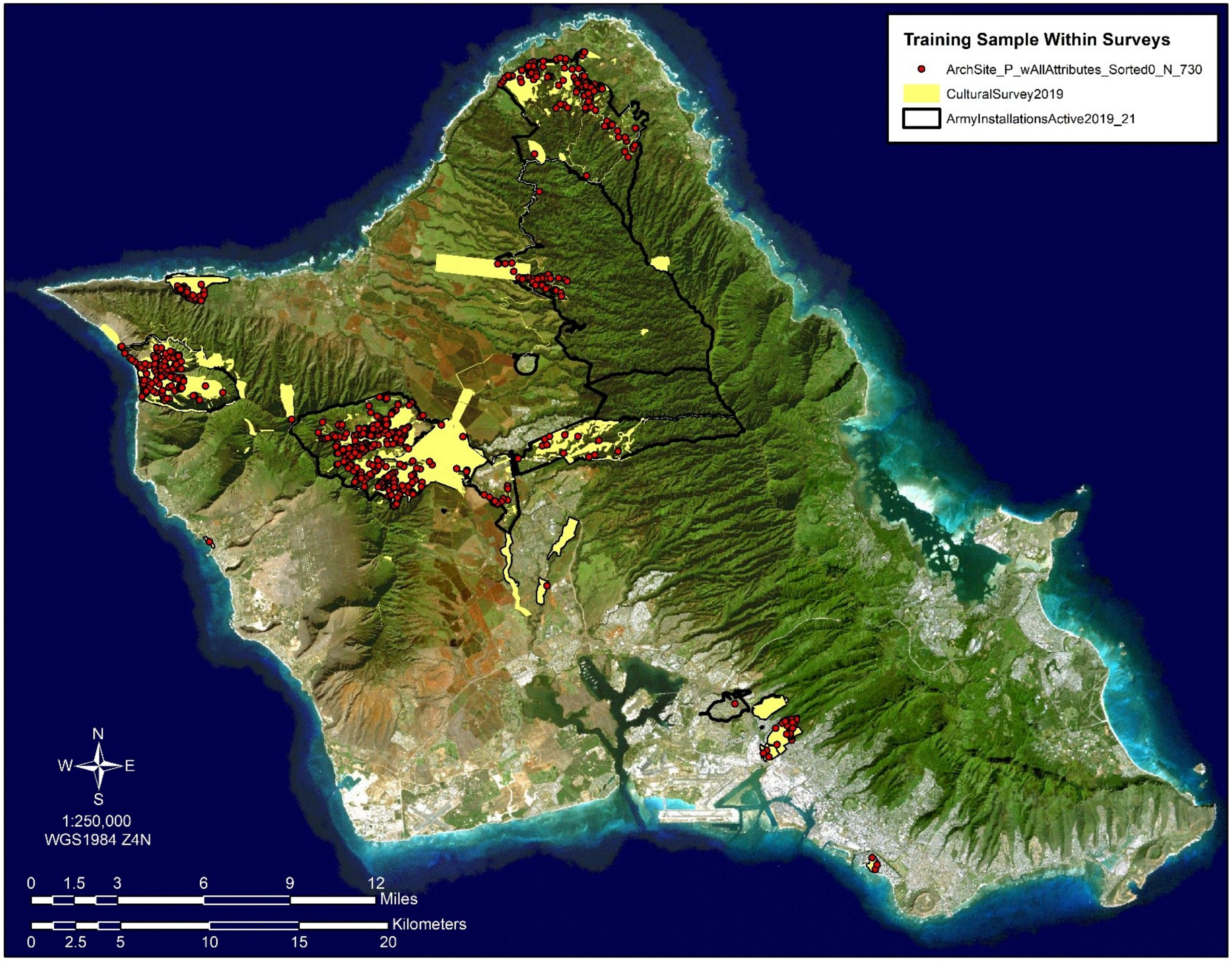

Training Sample Within Surveys
ArchSite_P_wAllAttributes_Sorted0_N_730
CultuarlSurvey2019
ArmyInstallationsActive2019_21
N
W
E
S
1:250,000
WGS1984 Z4N
0
1.5
3
6
9
12
Miles
Kilometers
0
2.5
5
10
15
20

Figure 1. Cultural Resource Survey Areas on Installations with Training Sample Sites.

secondary forests with an overstory consisting of Christmasberry (*Schinus terebinthifolius*), *Haole Koa (Leucaena)*, Albezia (*Falcataria moluccana*), ironwood (*Casuarina equisetifolia L*), Cook pine (Araucaria columnaris), Norfolk pine (*Araucaria heterophylla*) and *kukui* (*Aleurites moluccana*), likely a result early sandalwood deforestation and ranching.

Environmental Variables

Based on previously collected information, several environmental factors have been identified to provide general parameters to support identification of potential cultural resources on O‘ahu. Previous studies have noted elevation as a defining variable, and a majority of sites have been identified from sea level to approximately 1400 feet above mean sea level (masl) (Robins and Spear 2002). However, there have been some sites, such as trails and trail markers, found at higher elevations. Previous studies have mentioned rainfall gradients as a likely site limitation, but studies have not really investigated this. The degree of slope is also discussed in past studies, indicating sites range from 0 to 35-degree slopes with some exceptions such as caves and rock shelters. Prediction of trails, caves and rock shelters and significant landforms are identified through general topographic traits based on ridgelines and prominent peaks and rock outcrops that exhibit defined characteristics. Aspect, the primary orientation that a site faces, is not an attribute deeply examined in Hawai‘i archaeology so far, except in the case of specific heiau. Other regions have incorporated this into site identification studies, generally tied to cultural practices such as facing the rising sun. Proximity to fresh water, i.e., streams, is also noted but not well examined in Hawaiian archaeological studies despite the discussed relationship in many agricultural sites (irrigated and ponded) and habitation sites. Soil type is something that all survey reports discuss in general, but not well incorporated into expectations of survey results in advance. Land cover, or vegetation type, is usually discussed in survey plans, and reported results in terms of what is present in the survey area. This largely focuses on if it’s indigenous or invasive species, disturbance, or primary and secondary forests but not as a prediction indicator. Solar radiation, or insolation, is also not considered often in survey planning. Solar radiation is a derivative of atmospheric effects, latitude, seasonal sun shifts, and shadow effects, as well as aspect and elevation, some of which were examined separately in this study.

CHAPTER 3: BACKGROUND AND PREVIOUS STUDIES

Development of Site Location Modeling

Modeling, a simplified description, is a system or process used to represent something as an example or scaled imitation. Various attempts to create models to predict the likelihood of potential archeological and cultural resource site locations across a landscape have been created with limited success over the years. Predictive, probability, and suitability modeling have each been applied in other studies on the continental United States to anticipate locations of archaeological sites prior to ground exploration and are discussed in the following literature review. Although carrying different names, these models all consider desirable environmental qualities or known traits gleaned from known samples. Few advanced GIS studies have been found that specifically apply to island environments, including the Hawaiian Islands. Early settlement pattern analysis was based on intuitive assessment of other past site discoveries, some of which are discussed in the background section.

With the development of more sophisticated technologies, there are more capabilities for developing highly accurate topographic maps and integrating other techniques. High precision land surveys, and LiDAR, high-resolution aerial imagery and other remotely sensed techniques such as photogrammetry, can provide a better look at our environment. More precise environmental data is being collected and made available for general use with the accessibility of submeter accurate GPS and various monitoring sensors that can further enhance existing environmental knowledge, such as resource locations and rainfall.

Products such as digital elevation models (DEM) generated from LiDAR data rather than 7.5-minute quad maps are becoming more common and publicly accessible for inclusion in planning for archaeological surveys and land management.

Currently, commercial drones with cameras, one of the most cost-effective tools for small scale aerial photography, is not authorized on Army lands except for explicit permissions limited to military mission support.

These combined with more powerful computers and processing software, enable modeling that can be more complex and potentially automated. These products of environmental attributes and base imagery can be collated and overlaid to define probability zones for field verification.

Literature Review

A literature review was conducted to assess what past predicative and probability models for locating archaeological sites utilizing remotely sensed data have been used and what their strengths and weaknesses were. The most common remote sensing techniques identified as successful were also reviewed to assess their applicability to be incorporated into an Oʻahu model. With the recent advancements in remote sensing, this literature review largely focused on works after ca. 2000 unless there was a useful direct application.

In general, past probability and predictive models have not been found to be overly successful and have been unable to replace pedestrian surveys or adequately guide them. The past predictive success rates range between the 50 to low 80 percentages. Less than 70% was not considered able to justify extensive costs. Kvamme et al. (2006) describes the First Age during the mid-1980s as reflecting many obstacles that prevented models from being developed, including lack of computer technology. The use of predictive models waned in the 1980s and 1990s because of these low returns. Probability models are coming into a Second Age (2006:4) with the advancement, increased resolution quality, and increased availability of today's remote sensing techniques. This Second Age has advanced by the application of univariate and multivariate statistics, applications and development of GIS, and recognition of archaeological sampling bias. As techniques are refined and more commonly utilized more for other applications, there are more available data at a cheaper price (or free). Today, it is not always necessary to commission specific remotely sensed data for specific areas as many resources are publicly available.

Archaeological Site Location Prediction and Probability Modeling Overview

Predictive and probability models are employed to anticipate the likelihood of sites to be present in unexplored areas based on certain environmental attributes defined by a sample of known sites. Published studies have been limited in success rates and usability and few explicitly define their model methods. Ford et al (2009) utilized the Environmental Systems

Research Institute (ESRI) ArcSDM (species distribution model) extension that uses a Bayesian statistical method to weigh different environmental input layers such as soil fertility, soil drainage, slope and distance to watercourse, and then provided a breakdown and ranking of what is most influential. This method was used to predict settlement patterns of Maya civilization in Belize which yielded an 82% success rate overall in identifying site locations in the high probability areas. Duncan and Beckman (2000) attempted to develop and apply a reliable model using a flexible predictive GIS model using easily obtainable and consistent data that could be applied to almost any geographical location. Their study area was in West Virginia and Pennsylvania and they constructed a model within the GIS using a combination of inductive reasoning, statistical analysis, and spatial analysis of the raster module to produce an estimation of potential for site location in the high 70th percentile. The most successful predictive model discussed in Dalla Bona (2000) was applied in Ontario, Canada that was designed for forested site management. This model yielded 84% successful site potential with a deductive model using weighted values for basic environmental data such as elevation, slope, aspect and distance to water.

Recent works with various remote sensing technologies including converting LiDAR and Interferometric Synthetic Aperture Radar (IfSAR) into DEMs have been focused in Europe and the mainland United States. Doneus and Briese (2006) reported on the usability of waveform laser scanning in Austria. Their research discusses problems refining models by filtering to remove "blunders" and vegetation but not archeological features, reasoning that their appearance in the point cloud is very similar to natural and recent features. In Armenia, Parmegianai et al. (2003) compared DEMs from draping European Space Agency (ERS) Synthetic Aperture Radar (SAR) satellite imagery and topographic maps to assess morphological differences. This method presented problems with spatial resolution and precision resulting from shadow, foreshortening, viewing geometry of high relief areas and layover resulting in inaccurate placement of known sites on slopes. In another study, Menze et al. (2006) used Shuttle Radar Topographic Mission (SRTM) 3 arc-second terrain model for a virtual survey of *tells* (a prehistoric settlement mound) in the Near and Middle East. The resolution of the SRTM allowed for spatial observation of the larger features across the landscape, but the smaller ones were more difficult to distinguish with a semi-automatic detection strategy. Specifically, related but not applied to archaeology, Sanders (2007) had success comparing LiDAR, IfSAR, SRTM, and United States Geological Survey (USGS) National Elevation Dataset (NED) for modeling flood episodes. Menze claimed that LiDAR was the best terrain source with good horizontal and vertical accuracy achieving near bare earth or DTM at ~0.1 meter. However, the IfSAR and SRTM DEMs had good horizontal accuracy but more limited vertical accuracy, thus generating a surface model (DSM) with speckling that affected flood modeling. The NED was noted to be very smooth which is beneficial for flood simulation, but it was also limited vertically and produced a less accurate DSM and significantly underestimated the flood potential. These examples point to LiDAR as the best resolution to attempt bare earth models since it can penetrate the vegetation canopy more than other remote sensing techniques.

A site distribution study utilizing Fort Campbell Military Reservation, Kentucky-Tennessee site data, focused on statistical logistic regression and spatial analysis in GIS (Mills 2010). The statistical approach however was found unable to significantly quantify the relationship between archaeological sites and environmental variables. The site sample of 330 sites was also deemed too small, especially once split by Occupation period.

Archaeological Site Location Prediction in Hawai‘i

Research to date has unveiled few archaeological predictive or probability models that have been applied to Hawai´i.

As mentioned earlier, Anderson 1998 introduced a rudimentary predictive model for the USAG-HI. The “model” presented maps highlighting narrow bands in all stream bottoms as high probability, plateaus as moderate probability and steeper slopes, high elevations, and disturbed areas as low probability. As with most research presented in grey literature for Army archaeological studies, the model is largely conceptual yet remains true in a general sense, but is not overly successful in actual site prediction above basic logic.

Previous site location prediction modeling was largely focused on compilation and comparison of past survey results documents and gleaning out the common occurrences. Much of this was referred to as Settlement Pattern Analysis in archaeological reports rather than site prediction. These were largely discussions that reviewed desirable attributes. Anderson and Williams 1998 refers to and summarizes from Robins and Spear 1996:

> …lower elevation zones between 700-1000’, concentrations of permanent residences would be expected around major streams with associated religious structures. Irrigated agriculture and travel routes would also occur in these lower elevation zones. In moderate elevation zones between 1000-2000’, scattered temporary and permanent residences and heiau would be expected along streams, especially stream confluences. Irrigated and non-irrigated agriculture would be expected in these areas. In higher elevation zones between 2000-4000’, temporary habitation sites would be expected along travel routes. Some irrigated and non-irrigated agriculture would be expected in the lower elevations of these zones. Areas of legendary significance also occur at these higher elevations such as Mt. Ka‘ala and Kolekole Pass (Robins and Spear 1996:49).

Additionally, Green 1980 consolidated others researchers’ analysis of probably adaptations of early Polynesian settlers to the Hawaiian environments. Green’s compilation suggests initial population of wetter, more fertile areas on the coast first, where marine and terrestrial resources were accessible, then expanded along to other coastal environments, and then into the uplands.

Dega and McGerty 2002 also generalized site location expectations based on previous research. In a section on “Site Predictive Models”, they review several site types, such as irrigated and dryland agriculture, permanent and temporary habitation, ceremonial sites and trails, and the dominant commonalities of the setting and association with other site types, and focus on topographic and geomorphic zones and their different utilization. Valley shapes, nearness to water and general soil types and elevations per site type are discussed. The model basically described all agriculture occurring below 1100’ with irrigated on lower and flatter terrain than dryland, and habitation sites below 1000’ near agricultural sites, and trails that would connect large site complexes.

This type of modeling is generalized and only discussed in text rather than application. While some valuable indicators and environmental traits are brought to light, there were few correlations with other environmental attributes. Neither report presents useful maps indicating these high probability areas for planning. There are some priority and probability maps presented in Anderson and Williams 1998, but those are focused on prioritizing areas for survey because of training impacts or other threats to sites and management or probability zones for sites that are focused only as a narrow buffer on streams.

Williams 2004 conducted a study at Makua Military Reservation (MMR) to evaluate the potential usefulness of remotely sensed multispectral imagery against compilation of various archived aerial photographs and maps to locate potential cultural resources. The study used Thematic Mapper Simulator (TMS) spectral data collected from high altitude aircraft in 1985. The imagery collection exhibited cloud coverage in several areas therefore a single flight line (Line 25, Run 1) which covered most of the valley was used rather than a mosaicked radiometrically normalized image. Based on this imagery, the outcome of the comparison study determined that the compiled archival information was more applicable than TMS imagery. TMS was found to be of limited use for identifying potential cultural resources due to its low resolution. Williams noted that anomalies and irregularities could not be ground truthed to validate the spectral signatures due to the unexploded ordnance (UXO) hazards and the low relief of structural cultural remains at MMR could have been difficult to interpret. At this time, processing and analysis of the TMS data was four times the cost of archival data research.

Lessons Learned by Past Studies

For CRM, it is more cost effective to avoid areas with even a relatively low probability of site occurrence for project planning; however, this results in a large amount of area to be in question for development (Connolly and Lake 2006). A model that predicts a high percentage of sites over a large area would be helpful for large scale avoidance, but may not accurately reflect lower percentages of probability over much smaller percentages of land when attempting to ground truth site locations (2006). Likely, several variations in scale may need to be considered for cultural resources management by targeting high conflict areas first.

Connolly and Lake (2006) recommends statistical logistic regression analysis of the data over linear regression, stating it is more suitable to predictive modeling because it 1) can use a combination of differently scaled variables and 2) seeks to fit an “S”-shaped probability curve that allows for site presence to easily switch from high to low probability.

Ford (2009) used ArcSDM weights-of-evidence module to classify factors. This module, based on Bayesian statistics, computes a table of class weights, contrast, and an estimate of the variance of the contrast for significance testing. It also permits trial and error testing of class and layer combinations, layer independence, category aggregation, and statistical fitting of the model. Using ArcSDM would eliminate the researchers subjective weighting bias since it is generated within the module. Greater insight to site distribution across a landscape might enable augmenting the environmentally driven models with the concept of cultural entity (Lock 2006).

To consider the landscape as it was then, not necessarily how it is now, is a difficult endeavor. Generally, it is assumed in Hawai‘i that inland occupations did not take place until 1100 years ago so erosional processes in the uplands would still be noticeable, especially following the deforestation from sandalwood harvesting. For estimating locations of buried cultural resources, Ziedler (2001) discusses the Channel-Hillslope Integrated Landscape Development (CHILD) model that simulates long-term landscape erosion and deposition information and incorporates the varying storm duration and intensity to analyze geomorphic impacts of sedimentation to assess potential for buried cultural resources. There has been no subsurface shovel testing conducted on Army lands targeting possible buried cultural deposits. All testing has been conducted at sites with surface indictors. The CHILD model may provide some insight if it does not require known site data to predict areas of geomorphological sedimentation that could have buried deposits.

Processional models are relatively concise methodologically and reproducible within a GIS context (through buffering, overlay, and statistical tests of association) (Lock 2006). Lock also discusses the graphical intuitive approach (GIA) that might consider the type of site and how it would be used; would the site pull people to it (ie. *heiau*, trails, and agricultural plots) or repel them (kapu), or have limited use (severe slopes)? Considering this in weighting variables or in nearest neighbor analysis could also be applied (2006:). Augmenting environmentally driven models with the concept of cultural entity might also enable a greater understanding of the cultural landscape (2006).

Archaeological Background

Initial occupation of the Hawaiian Archipelago is currently believed to have occurred at the end of the first millennium and between A.D. 780-1119 with a mode of A.D. 960 (Bayman and Dye 2013) based on calibrated dates from floral and bone materials. There is still debate over which Polynesian settlers arrived first and occupied which Hawaiian Islands before others. Traditional Hawaiian archaeological sites range from constructed dry-stone surface structures (e.g., walled enclosures, terraces, platforms, fish ponds, and agricultural mounds), to fire hearths, food and tool manufacture middens, taro ponds, and excavated earth ovens (*imu*) below ground, plus rock art. Many of these types of surface structures are still visible on the terrain where historic and modern civilization has not overcome them.

Captain James Cook, sailed past O‘ahu in 1778 just one year before landing on Hawai‘i Island to his demise, described O‘ahu as having rich, fertile, and cultivated valleys. However, thirteen years later Captain George Vancouver noted that agriculture and the population itself were diminishing (Handy and Handy 1991). This seemingly abrupt change could have been the result of introduced disease, inter-island warfare, and seasonal changes in residency (Anderson and Williams 1998). Following this initial foreign contact and traditional religious upheaval, land uses changed significantly to refocus indigenous work efforts on harvesting sandalwood for trade to China, which was followed by sustained foreign settlement and the purchase of fee simple land, the development of cattle and sheep ranching in the 1850s, cultivation of sugarcane in the 1890s, and pineapple in the 1910s.

Settlement and Agricultural Intensification Patterns on O‘ahu

In Hawai‘i, many coastal areas and agriculturally productive areas have been built up in the last 230 years since Captain Cook made contact with Native Hawaiians in 1778 and began the influx of foreigners that dramatically changed traditional land use, subsistence patterns, settlement patterns, and construction styles. General settlement on O‘ahu has been shown to originate from three main settlement centers according to radiocarbon dates recovered from archaeological excavations. From an O‘ahu island perspective, settlement patterns originated from the mokus of ‘Ewa on the leeward coast near current Waianae, Kona near current Honolulu, and Ko‘olaupoko near Waimanalo (Cordy 2002). Settlement patterns on the island are thus examined based on environmental variations, subsistence development, population changes, political and social organizations. Archaeological research infers that habitation settlements originated on the coast and resource procurement occurred in the uplands in daily or short-term excursions. As habitation settlements grew and resource demands expanded, more permanent habitations spread further into the uplands and agricultural intensification increased as well.

Wai´anae Uka was considered the seat of power for O´ahu from ca. 1100 until Chief Mā´ilikūkahi moved it into Waikiki ca. 1550 (Desilets 2011). Ethnographic information indicates large populations living in central O´ahu to support the power seat. Despite extensive agricultural and urbanization disturbance in the lower saddle of Central O´ahu, the number of Hawaiian sites known today in what was Wai´anae Uka appears hugely deficient to support the estimated large populations. Much of the uplands were once denuded by sandalwood harvesting in the early 1800s, which sometimes include burning. This harvesting practice denuded far more forested land than was just disturbed by carts, dragging harvested brush and creating staging areas from which to move the harvested wood to the harbors to sell. These areas remained sparsely vegetated throughout the following ranching and grazing era.

Ranching would have initially had a limited impact on many traditional upland Hawaiian feature types beyond grassland settings. Likely, the largest impacts would have resulted from establishing remote stations away from the main ranching complex and it has been documented that some ranchers dismantled some features to build other structures such as livestock corrals (Dixon et al. 2004) or dams (Henry 1992). Intensive agriculture for pineapple began in the late 1800s and was focused primarily on the lower elevation plateaus. A 1916 topographic map with hand-drawn annotations show most of the lower elevations of the modern training ranges cultivated by pineapple and includes some substantial historic irrigation ditches. According to this land use scenario, upland military lands have potential to still contain traditional Hawaiian site types including habitation sites, agricultural complexes, religious centers, and other less visible features such as trails, *ahu* (shrines), utilized caves and rock shelters, sacred places, and other special use sites and buried deposits that may not have visible surface features.

For the purpose of modeling, the precise timeframe of the expansion is irrelevant but the attributes tied to the chosen areas are important. Habitation and agricultural sites are being targeted in this study as having the most specific targetable variables. Other site types have fewer overlapping attributes and may require different modeling attributes, but may also be selected.

Cultural Resources Management on USAG-HI Lands

Previous pedestrian archaeological surveys conducted on O‘ahu USAG-HI lands have identified traditional Hawaiian, early ranching, agricultural features from pineapple and sugarcane cultivation, and early military sites. These sites consist of general feature types such as terraces, enclosures, platforms, mounds, walls, and wet (*loi*) or dry (*kula*) agricultural fields. Today, many lands have been heavily modified by historic sandalwood harvesting, ranching, intensive agriculture, and increased populations while modern urbanization has obscured, impacted, or destroyed archaeological sites. Much of the USAG-HI lands are currently designated as training areas and are largely forested. These middle and upper elevation lands are less disturbed by urbanization.

Survey in dense vegetation limits ground surface visibility for artifacts and low-lying features. Large features are easily obscured in dense grasses and vines but more likely to be observed. Current reviews of older pedestrian surveys are being found to have used intervals more widely spaced. Re-survey in many areas are identifying totally new sites and expanded known sites. Early fieldwork focused on large structures, either overlooking low agricultural complexes because of the focus of the survey or perhaps from lack of recognition of the subtle modifications. Many early studies were reconnaissance level, which focused on presence or absence of sites to identify the potential in a sample area. In the mid-1990s, surveys became more comprehensive and there seems to have been a greater awareness of more site types, paying attention to less developed stone construction rough sites and therefore finding features that were previously overlooked. From the mid-1990s, surveys expanded their research scope to inventory smaller and single features. The spacing of the individual survey team members was also reduced to attempt greater than 80% survey coverage. Today’s team members are spaced 5 to 15 meters apart depending upon the terrain and density of vegetation, and the level of awareness for such feature types is more common judging by current research results in the islands. Even for some surveys in the 2000s, some 30-meter survey widths in open terrain such as open lava fields on the Island of Hawai´i are not considered adequate now as cultural features continue to be found in supposedly surveyed areas. At the time there seemed to be adequate visibility of the open ground surface and between survey team members. However, upon conducting the next phase of detailed site recording, the identified sites expanded and some new sites were found within the undulating terrain that would have been between transect lines.

Some of these inconsistencies are subjective based on the project directors' judgment and guidance, others are tied to the experience of the individual's knowledge to recognize a cultural feature and some are due to changing environments. Wild fires and prescribed burns of land parcels reveal new cultural features as do erosion from long term construction projects and natural processes. For example, at Schofield Barracks Military Reservation, prescribed fires were utilized to expose the ground surface to facilitate clearance of UXO, construction and cultural surveys. Many sites were found on the exposed landscape previously covered with dense tropical forest. After several years of annual prescribed fires, soil and wind erosion exposed additional features in previously surveyed areas.

CHAPTER 4: PROJECT METHODOLOGY

Based on the previous studies and considerations of local Hawaiʻi setting, indicative environmental variables have been considered such as aspect, elevation, rainfall, slope, soil type, hydrology, solar radiation and land cover. All environmental attribute data were appended to known site and randomly sampled point data using ArcGIS 10.4 or 10.5. The composite feature class data was then exported to an excel spreadsheet and analyzed in either StatView 5, an Analytics Software and Solutions (SAS) Institute statistical program, or R 3.5.3 statistical program to identify significant attributes. Matrix Laboratory (MATLAB), a multi-paradigm numerical computing environment, was then used to model the significant attributes and export a probability map of Oʻahu island for location expectations of cultural resources.

This research and data access are in agreement with the USAG-HI under conditions of employment by the author at the Directorate of Public Works, Environmental Division. Applicable predictive suitability modeling results are intended to be utilized for the USAG-HI cultural resources management and included in ICRMP.

Cultural Resources Data: Training and Test Samples

The Army Environmental Command (AEC) created and requires management of all army geographical information in a Spatial Data Standards for Facilities, Infrastructure, and Environment (SDSFIE) geodatabase (GDB). A GDB is a collection of feature datasets, feature classes, object classes, and relationship classes to manage a seamless continuous spatial representation. For coordination with the nation-wide army SDSFIE program, the data is maintained in World Geodetic System 1984 (WGS84), Zone 4 North.

The USAG-HI SDSFIE GIS site database was gleaned for less accurate location information and reduced to 730 sites to serve as the training sample. The validation test sample consists of 89 sites compiled from three recent surveys conducted in the Kahuku Training Area (KTA) during the development of this effort. These site data were recorded with GPS at 1-meter or better accuracy collected by in-house staff or provided under external archaeological contracts.

Training Sample

The site data were examined and reduced from the total sample of 1009 sites to 730 as shown in Figure 1. Sites were removed from the sample if they were evaluated to be untrustworthy for some reason or deemed not to be a human-derived feature or site. Untrustworthy sites include sites with locations from surveys older than 2000 with questionable accuracy of locational data or those known to be manually digitized (e.g., Whitehead via personal communication that Universal Transverse Mercator coordinates (UTM) derived from a template on a map were then digitized in GIS). Additionally, points representing linear data were removed, such as a point representing a trail. Structures, such as buildings and bridges were also removed; however, remnant military structures such as pillboxes and gun turrets were retained.

For the purposes of training sample for this study, only the point data were used. Most of the polygon data were found to be arbitrarily buffered off a point or inaccurately plotted. The polygon data are in the process of being updated. The site points (N=730) were coded as “1” to indicate the presence of a cultural resource.

Test Sample

During the writing of this project, three surveys were funded and completed. The results of the survey findings serve as the validation test sample. Three surveys were conducted at the KTA that were within the environmental setting parameters as the training data. While within the environmental attribute ranges, the surveyed areas are contiguous and do not cover a wide variation of the parameters. These surveys were conducted based on training needs prior to the model creation. Combined these surveys identified 89 new cultural resources. The test sample sites were also coded as “1” for validation testing at the end.

Creation of Random Sample

In order to obtain an unbiased sample, a random sample of 730 points was created from within the same survey polygons as the known training site sample (Figure 2). The random points were generated in ArcGIS, using the Random Point Generator tool under Data Management, Sampling. The data was created in WGS84 to be used with the site point data.

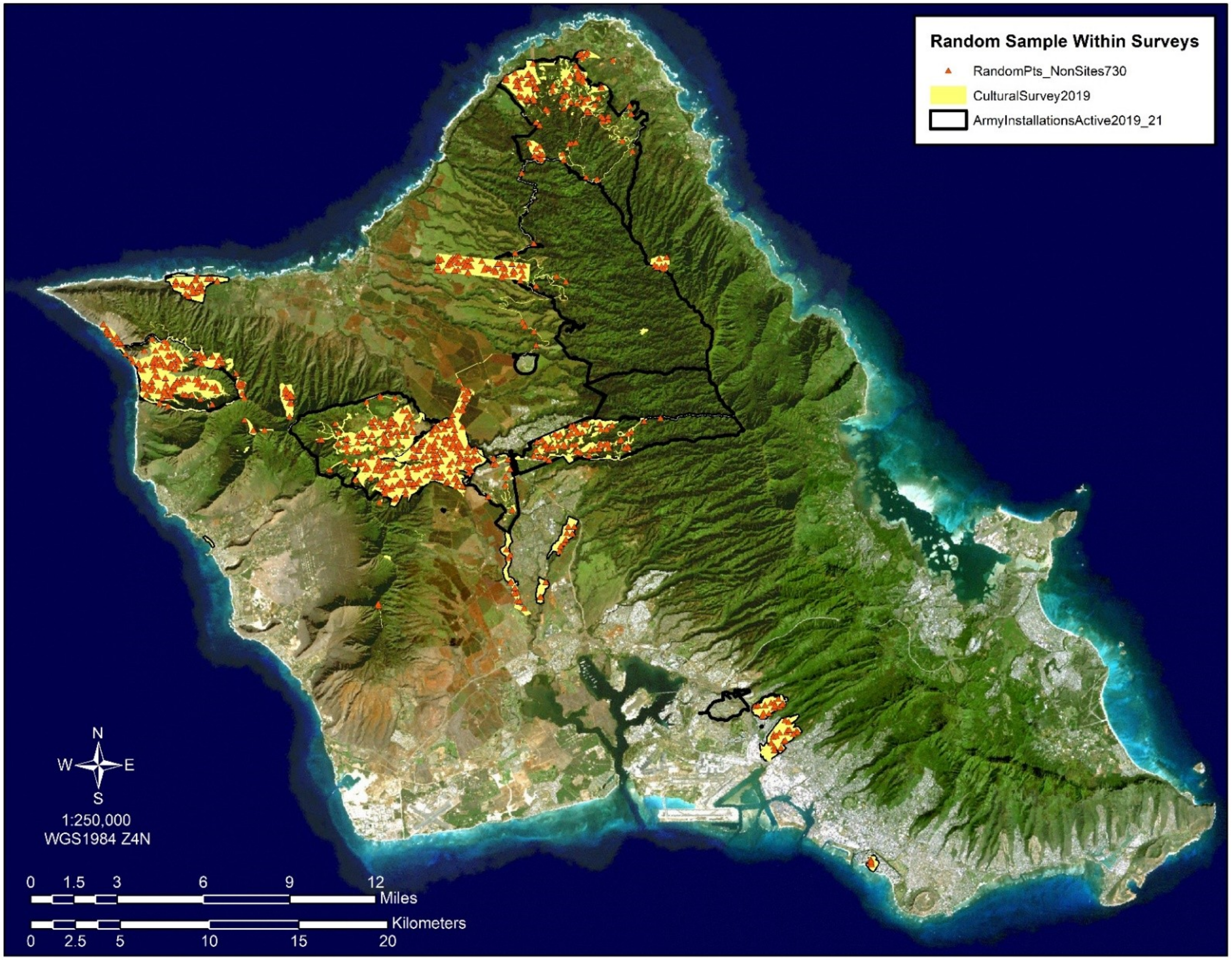


Figure 2. Cultural Resource Survey Areas on Installations with Random / Non-Site Points.

A separate random sample of 89 points was created from within the survey areas that the known sites of the training sample were discovered in.

All point locations were appended with environmental attribute data by the same process according to the processes described in the following section. Four separate feature classes were created for the 730 known training and 89 test site points and for the corresponding 730 and 89 randomly generated points. The 730 randomly generated points were combined with the 730 training sample points into a common spreadsheet for the second phase of statistical analysis.

Selection of Environmental Variables

Environmental data was compiled from publicly accessible sources. The data were manipulated in a GIS platform, running ESRI ArcGIS. Environmental attributes were extracted from raster and vector data and the cell values were appended to corresponding point data feature classes. Since the cell values were represented by a single point, only one 10m by 10m raster grid cell represents each site, along with the characteristics of each environmental attribute of that corresponding raster cell. Any site larger than 10m, which was a large portion of the sample set, was not fully represented.

The corresponding environmental data for each training and random sample points were appended into one master feature class as a "description" of the point location. Once all attributes were converted to numerical or coded data, they were exported as an Excel table, imported into StatView, R, and later MATLAB, for various stages of analysis. Data for aspect, elevation, rainfall, slope, and distance to fresh water were utilized as discrete real numbers. Solar radiation data existed in zones and was utilized as a rank based on the lowest measured value of the pre-defined zone. Soil character and land cover data existed in coded values and were used as categorical data.

Aspect

Aspect is the cardinal, or compass, direction the land surface faces and is measured in degrees from magnetic North. The orientation of the surface plane on which a site is situated was examined for its commonality with other known sites. Aspect is strongly tied to solar radiation, which is discussed below. Aspect may be significant as an indicator of sun exposure or directional orientation that may be tied to a cultural practice, such as the rising or setting sun.

Aspect was derived from an existing DEM downloaded from the School of Ocean and Earth Technology (SOEST), University of Hawaiʻi at Mānoa website. The DEM was created by the National Oceanic and Atmospheric Administration (NOAA), National Ocean Service (NOS) is from 7.5-minute quads (i.e., 10-meter resolution) in North American Datum 1983 (NAD83) UTM, Zone 4 North and published in 2007.

A new aspect raster was generated automatically by a comparison algorithm of the orientation of the eight surrounding cells through the ArcGIS Spatial Analyst Extract tool. From this new raster, cell values for each point were extracted from the raster and appended to the master feature classes.

Elevation

Elevation is the height of a given item. The elevation of sites were evaluated to identify common or uncommon occurrence in a given height or range of height. Sites have been previously identified to occur between sea level and 1000' for habitation or temporary sites and up to 1400'for agricultural sites (Robins and Spear 2002). Most of Oʻahu is accessible by human activities, however lower elevations that correspond to sizable relatively level terrain and accessibility to multiple resources trend towards more permanent habitation areas.

Elevation was derived from the same 10-meter NOAA DEM downloaded from the SOEST public website. Using 3D Analyst Functional Surface tool, elevation values were extracted from the DEM for each cell correlating with a point and appended to the point feature class. As with aspect, many sites are larger than the 10m grid, therefore the site is not fully characterized by this singular cell value.

Rainfall

Rainfall is the amount of rain received in a given area. This can be measured by day, month or annually. Rainfall is an important resource for growing crops or native vegetation and providing fresh drinking water if fresh water streams are not available. Rainfall can also contribute to the coolness or humidity of an area and recharge the aquifer sometimes accessible through caves or at low tide seeps on the coast.

Recently compiled rainfall data were downloaded from the online Rainfall Atlas of Hawaiʻi (Giambelluca et al. 2013). The raster data for annual rainfall in inches for Oʻahu were used.

Rainfall cell values were extracted using ArcGIS Spatial Analyst tool for each site point and appended to the point feature class for export into StatView.

Slope

Slope is the angle of the surface plane, calculated as rise over run and is represented in degrees or percentage. A slope of 45-degrees is equal to 100%, or a 1:1 slope ratio, which is quite a steep angle for most types of human activity (with the possible exceptions such as caves, rock shelters or hōlua slides). Human occupation generally takes place in areas of fairly even ground whether it is a living area, work area or agricultural area. Agricultural complexes might reflect a greater slope change but will vary on slope stability per soil type. From a geologic perspective on soil stability, the angle of repose or failure for clay soils is between 25 and 40 degrees for dry clay and 15 degrees for wet clay (Clover 1995). The angle of repose is the angle at which a soil type will fail, resulting in landslide or slump given basic conditions. High clay content soils will begin to slump and accumulate slumpage at the base of slopes exceeding 33.5° (Clover 1995).

Similar to aspect, a new slope raster was also derived from the 10-meter NOAA DEM for each site using Spatial Analyst, Extraction tool. From the new raster, cell values of slope were extracted for each point location and appended to the feature class for export.

Distance to a Stream: Fresh Water

This is the distance from an activity area to fresh water. Accessibility to fresh water for either agricultural or habitation use is a relationship that has been discussed in many past archaeological research studies. The optimal distance between an activity area and fresh water may vary per site type. This may also correlate to slope and elevation as an activity area may be very close as the crow flies, but is separated by a steep cliff.

Shapefile line data were downloaded from the City and County of Honolulu, Honolulu Land Information System (HoLIS) in NAD83 HARN UTM. These line shapefiles were available as perennial, intermittent or “all streams” (both perennial and intermittent) and measured in meters. Linear data for “all streams” were utilized and the distance from streams (fresh water) was calculated using the ArcGIS proximity tool for each site point. The discrete distance away from the stream line was appended to the point feature classes for export.

Solar Radiation

Solar radiation is the amount of incoming solar insolation (direct and diffuse) calculated for each site location. The amount of sun and heat generated at a given location could be an indicator of site location selection for growing crops or living areas. The amount of sunlight agricultural sites receives per day affects its productivity correlating with the plant type. In general, the insolation rate increases with elevation.

Solar radiation is a product generated in the ArcGIS Spatial Analyst toolbox derived from the NOAA 10-meter DEM. The data exported as six polygon zones measured in increments of 50 kilowatt hours per square meter (WH/m^2) of solar radiation. Each point was correlated to an increment and appended to the point feature class.

Soil Type

Soil, the type of ground above bedrock, can be categorized in a number of ways. Categories included Major Land Resource Areas (MLRA), Common Resource Area (CRA), map key unit (MUKEY), horizons, order, suborder, Great Group, Subgroup, Family and series. The lower the category, the more units are represented. The type of soil a site is situated upon may contribute or drive the function of the activity. The type of ground was examined to see if it was an indicator of functional use. Past research discusses certain soil types to be more desirable for agriculture or habitation or other uses. Additional soil survey categories can be explored for correlation to certain soil types or fertility.

Soil data were downloaded from the United States Department of Agriculture, National Soil survey Center (USDA/NRCS). These data were published in 2006 and are projected in WGS84. The dataset, largely polygons, is derived from the Soil Survey Geographic (SSURGO) Database. A maze of data can be linked to the raster. The six larger units from the MLRA dataset were utilized as a digestible data size of six units in this phase of analysis. The map key unit (MUKEY) classifications were explored, but found to be an overwhelming number of units and failed the initial significance test. The assigned numerical MLRA SSURGO codes describing dominate physical characteristics were used. Each site point was correlated to a soil unit using the Spatial Analyst Extraction tool and appended to the common feature class.

Land Cover

Land cover describes what is on the ground surface. Land cover classifications are broken down into Levels. Level I includes large units such as water, forest, rangeland, agricultural land, etc. Level II is broken down into subunits. Forestland, for instance, divides into deciduous forest land, evergreen forest land, and mixed forest land. Not all subgroups are represented on Oʻahu. These Level II subgroups were used in the analysis and are discussed below.

Land cover data were downloaded from the Office for Coastal Management in NAD83 UTM. Two datasets were available; lulc (land use/land cover) and ccap (Coastal Change Analysis Program), representing many land use and forest coverage attributes. Oʻahu data were selected and separated from the rest of the island chain information. While it is acknowledged that land cover changes over time, the ccap data derived from high-resolution QuickBird multispectral imagery in 2005 were utilized for this study. Oahu has 11 out of 23ccap land cover types occur that are coded numerically. The ccap class codes were correlated to the points using the Spatial Analyst Extraction tool and appended to the feature classes.

Data Processing

The four composite point feature classes were exported in Microsoft Excel for editing. Any remaining null values were populated as appropriate and any unnecessary feature class attribute fields were deleted as part of the data preparation for import into StatView, R and MATLAB. Initially, the 730 known site training points were examined to evaluate the sample dataset. For use in StatView and R programs, the Excel spreadsheets were exported as a tab delineated text file. The training site spreadsheet was later combined with the 730 randomly sampled dataset to conduct an unbiased binary logistic regression analysis to further describe the data and explain the relationship between the binary dependent predictor variable (Y) (presence/absence of a cultural resource) and one or more nominal, ordinal, interval or ratio-level independent response variables (X) (Table 1). The combined spreadsheet was analyzed in R and MATLAB. Both the test point feature classes were used for later model validation.

Table 1. Summary of Variables.

Variable	Variable Description	Variable Unit	Data Type
Predictor	Archaeological Site (1) or non-site/random sample (0)	1 Present, 0 absent	Binary Categorical
Response	Aspect	Degrees	Ratio
Response	Elevation (Z)	Meters	Ratio
Response	Rainfall	Inches	Ratio
Response	Slope	Degrees	Ratio
Response	Distance to Stream	Meters	Ratio
Response	Solar Radiation	Watt hours per meter squared (WH/m2)	Ordinal Categorical
Response	Soil Character (MLRA)	Numerical Code	Nominal Categorical
Response	Land Cover Class	Numerical Code	Nominal Categorical

For use in MATLAB, all data had to be in the same raster format. All attributes except for streams, were already in a raster format. Each raster was projected using Data Management, Projections and Transformations, Raster, Project Raster using a common registration point based on the centroid of the Oahu Island Outline feature class polygon. To run successfully, a new blank Map View had to be opened and the Environment Settings selected for the Processing Extent set to the Oahu Outline feature class and the Raster Analysis cell size set to 10 meters. Using “Snap Raster” option under the Environments Settings produced inconsistent columns and rows.

The HoLIS stream shapefile line required conversion to a raster. First the line was rasterized using the Conversion, To Raster, Polyline to Raster Toolbox, set to a 10m cell size to match with the other rasters. Second, the distance from the rasterized stream line was calculated for 10m grids using Spatial Analyst, Distance, Euclidean Distance. Finally, the raster was projected using the same registration point. The columns and rows were off by one using the same Oahu Outline polygon extent, so the final product elevation raster was used as the extent and resulted in a match. The end results included eight 10m-grid rasters with matched and overlapping columns (6574) and rows (5055).

Statistical Analysis

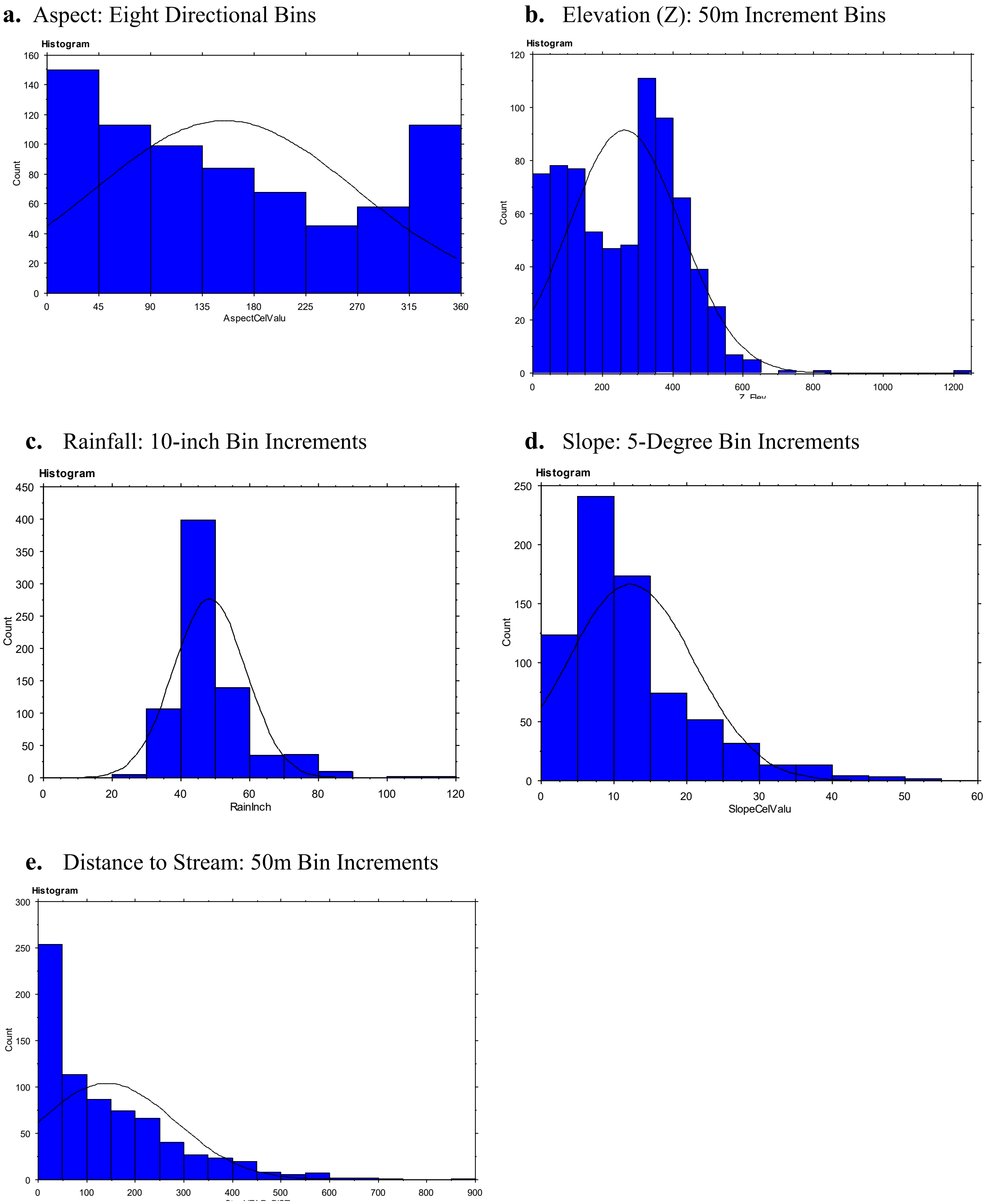
Descriptive Statistics

Analysis of the known 730 training sites was conducted in StatView to guide later model selection. The statistics within ArcGIS were limited to one data type at a time and were not exportable as a table for manipulation; therefore, it was not utilized in this analysis. Analysis of the descriptive statistics indicated the central tendencies were not tightly centered for any of the continuous ratio environmental attributes (Table 2) for the training dataset (N=730). Additionally, the standard deviation was always lower than the mean, confirming the non-normality and slight positive skew. Skewness for α(2) at 0.05 for a sample of 730 is 0.177 Skc.

Table 2. Descriptive Statistics for Environmental Variables for Sites (N=730).

Attribute Variable	Mean	Median	Mode	Std. Dev	Range	Skewness
Aspect	154.88	135.0	135.0	112.95	359	0.39
Elevation (Z)	261.63	281.57	1.00	159.21	1201.42	0.34
Rainfall	48.29	46.36	50.36	10.51	93.76	1.73
Slope	12.27	10.02	3.20	8.75	52.91	1.49
Distance to Stream	142.39	98.87	None	140.11	897.83	1.43

The analysis of the environmental attributes for sites indicated all data as non-normal, asymmetrical distributions with a positive skew as seen in Figures 3 and 4. The attributes were broken into bin units appropriate for each attribute to provide useful visual analysis. Aspect (a) was broken into 8 quadrants based on 45-degrees (°), elevation (b) and distance to stream (e) were displayed in 50m increments, rainfall (c) was broken in to 10-inch (") intervals, slope (d) was broken in to 5° increments, and the categorical attributes (Figure 4, a-c) were based on their codes. Non-normality is not uncommon or unexpected for environmental data. Given that the numerical and categorical data are not normally distributed, they cannot be analyzed with tests designed for normally distributed data and requires some degree of transformation.



a. Aspect: Eight Directional Bins

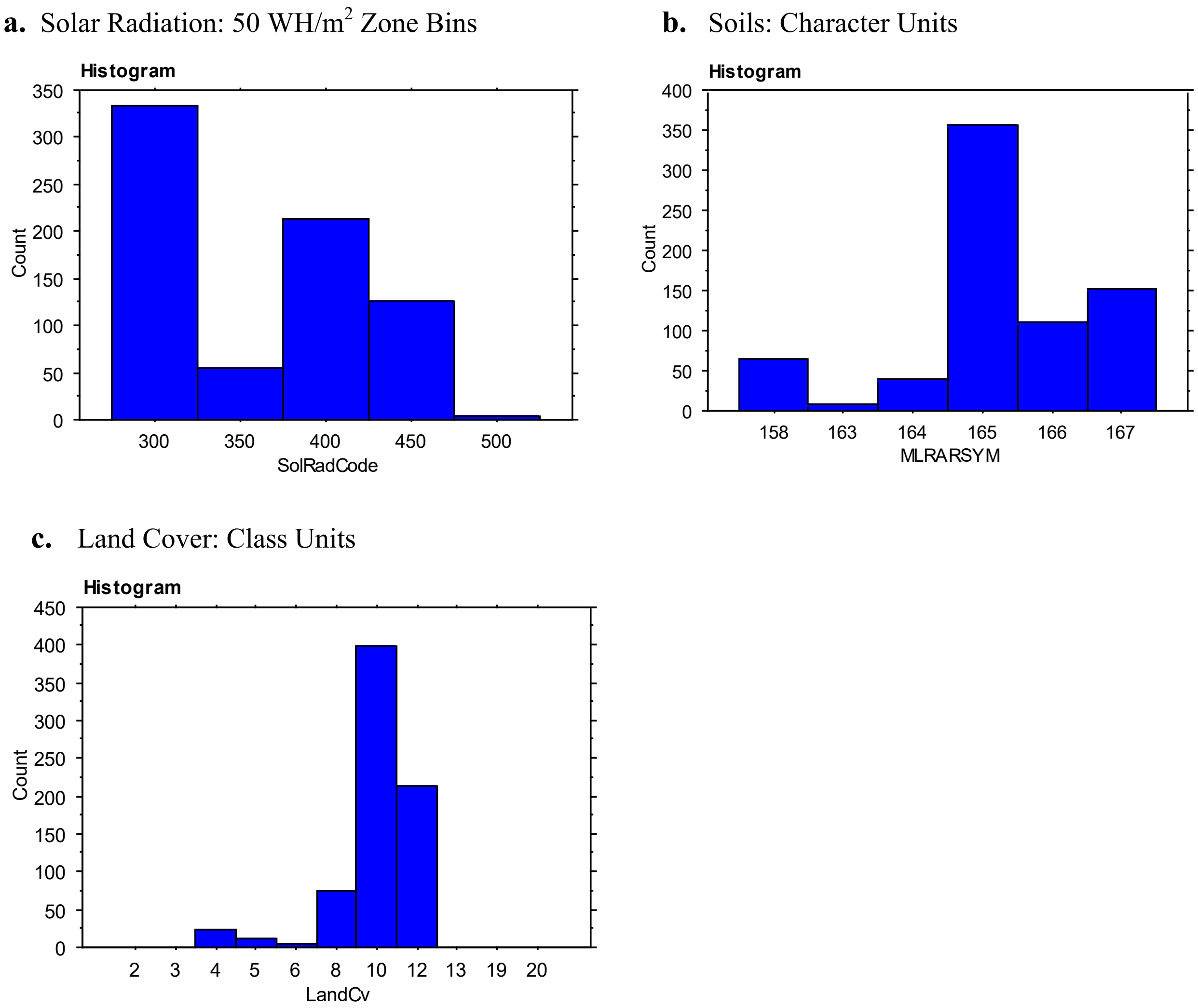
b. Elevation (Z): 50m Increment Bins

c. Rainfall: 10-inch Bin Increments

d. Slope: 5-Degree Bin Increments

e. Distance to Stream: 50m Bin Increments

Figure 3. Continuous Environmental Attribute Histograms (a-e) with Normal Curve Distribution (N=730).



a. Solar Radiation: 50 WH/m^2 Zone Bins

b. Soils: Character Units

c. Land Cover: Class Units

Figure 4. Categorical Environmental Attribute Histograms (a-c) (N=730).

Inferential Statistics

In order to obtain an unbiased sample for comparison, the 730 known site points were combined with 730 random sample points (total N=1460) that were generated in ArcGIS; each point was appended with the same environmental attribute data.

As non-normally distributed data, the environmental attributes were tested using non-parametric paired-means comparison for the continuous data and chi-squared test for the categorical data to determine if there is a statistically significant difference between the site and non-site/random sample. Both tests are applicable to non-normally distributed data.

In this case, the null hypothesis was that there is no difference between the known sites and random points, H_0: $\mu_1 - \mu_2 = 0$ (Moore et al. 2012). The standard p-value of 0.05 was used to determine the significance threshold. P-value is the chance that the result is true due to random variation. Commonly a p-value of 0.05 or less indicates that the result is significant (or that a 5% chance or less of this happening is just due to random variation). The five continuous environmental variables were analyzed independently using paired samples *t*-tests (Table 3) to determine if a difference between the sites and non-sites exists. Aspect, rainfall, and distance to a stream yielded a p-value of less than 0.05, which indicated a significant difference between the known sites and randomly generated locations. Elevation and slope did not yield a significant difference.

Table 3. Paired Means Comparison H_o=0 of Numerical Variables of Sites Verses Non-Sites.

Variable (unit)	Mean Diff.	DF	t-Value	P-Value
Aspect (°)	46.3	729	8.0	<.0001
Elevation (m)	-1.2	729	-1.4	0.89
Rainfall (“)	2.3	729	2.6	<.0082
Slope (°)	0.6	729	1.0	0.31
Distance to Stream (m)	91.1	729	10.4	<.0001

To determine the relationship between the sites and non-sites for the categorical environmental variables, they were independently analyzed using the chi-squared test (Table 4). The null hypothesis of the chi-squared test was that the two variables are independent and the alternate hypothesis was that they are related. For the Solar Radiation (SolRad) data, the p-value was less than 0.05; therefore, the null hypothesis was rejected and found to be significant, which indicated a difference between the sites and randomly selected non-sites. The significant differences for aspect, rainfall and distance to stream, indicates these variables may be useful in identifying site locations.

Table 4. Chi-Squared Tests for Categorical Variable.

Log Likelihood	DF	Chi-Square	P-Value
SolRad (WH/m2)	1	32.1	<.0001

The values for land cover and soil were excluded from the analysis since the numerical values are codes and not levels.

There is no statistical linear relationship with the environmental variables and the known sites or the random non-site points. A linear relationship is a graphical representation that results in a straight line between two separate variable datasets. The simplest representation is when variable X and variable Y increase at the same rate. Scattergrams are commonly used to demonstrate the strength of a linear relationship. In the case of this dataset, the known sites and the random non-sites are the dependent predictor Y variables, which are nominal, and the independent response environmental attributes are the continuous X variables. The scattergrams in Figures 5- 12 show the point graphs distinguishing the distribution of each variable, but does not describe any linear relationships between the dependent and independent variables. Because there are no linear relationships, correlations between variables cannot be made.

Figures 5-12 also show histograms illustrating each environmental variable distribution comparisons for sites (1) and random non-sites (0) along with the bivariate scattergram.

For aspect, all eight defined bin orientations were represented in both samples (Figure 5). For known sites occurred more concentrated between 315° to 135° with a peak between 315° to 45°, a northerly aspect and very low between 225° to 170°, a southwestern aspect. The random non-site distribution fell more consistent across the whole aspect range, but also showed a similar peak between 315° and 0° north. The absence of known sites oriented west may be a limitation of the available data, as USAG-HI does not hold much land that is oriented to the south and west and most of those areas have minor topography creating a more open aspect.

Visual dispersion of sites and non-sites was more varied for elevation than aspect. In Figure 6, viewed in 50m bins, known sites occurred concentrated below 500m with a slight bimodal frequency with lower peak between 0-200m at counts in the 70s and higher peak between 300-450m at 111 maximum count. The saddle between the bimodal peaks was near 50 sites. Site occurrence decreased rapidly from 66 sites from 400m to 5 sites at 650m and a few single sites up to 850m and one isolated site at 1202m. The non-sites occurred consistently in the 40's to 80's between 0 to 400m except for a spike of 178 points between 250-300m. From 400m, points decrease steadily from 29 points to four at 750m, then trickled up to 1250m.

Dispersion of known sites for rainfall peaked significantly between 45" and 50" per year at count of 398 (Figure 7) and showed a lower concentration in the low 100's between 30" to 50" per year with a few more sites receiving up to 90" per year. The non-sites showed a high count

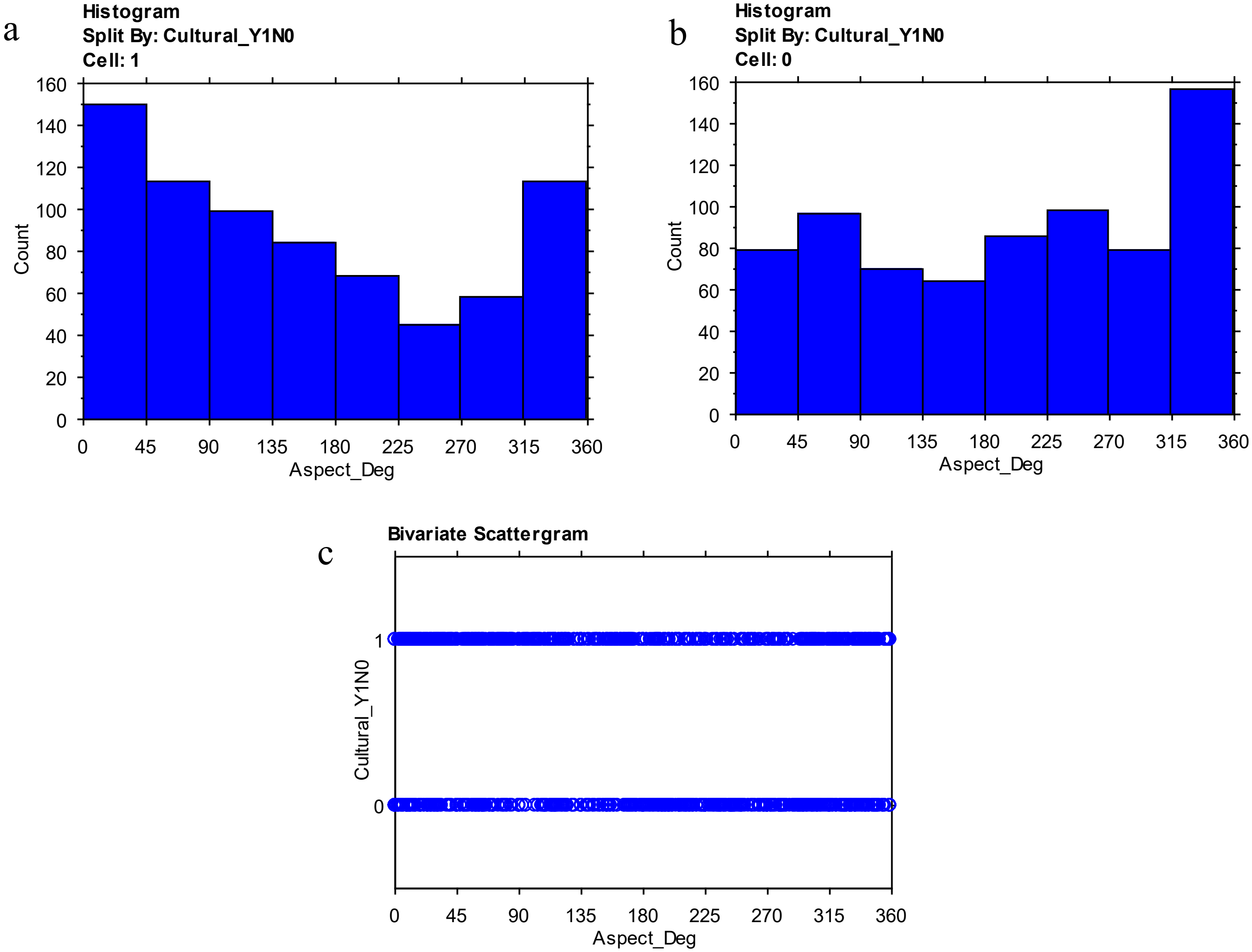


Figure 5. Aspect Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

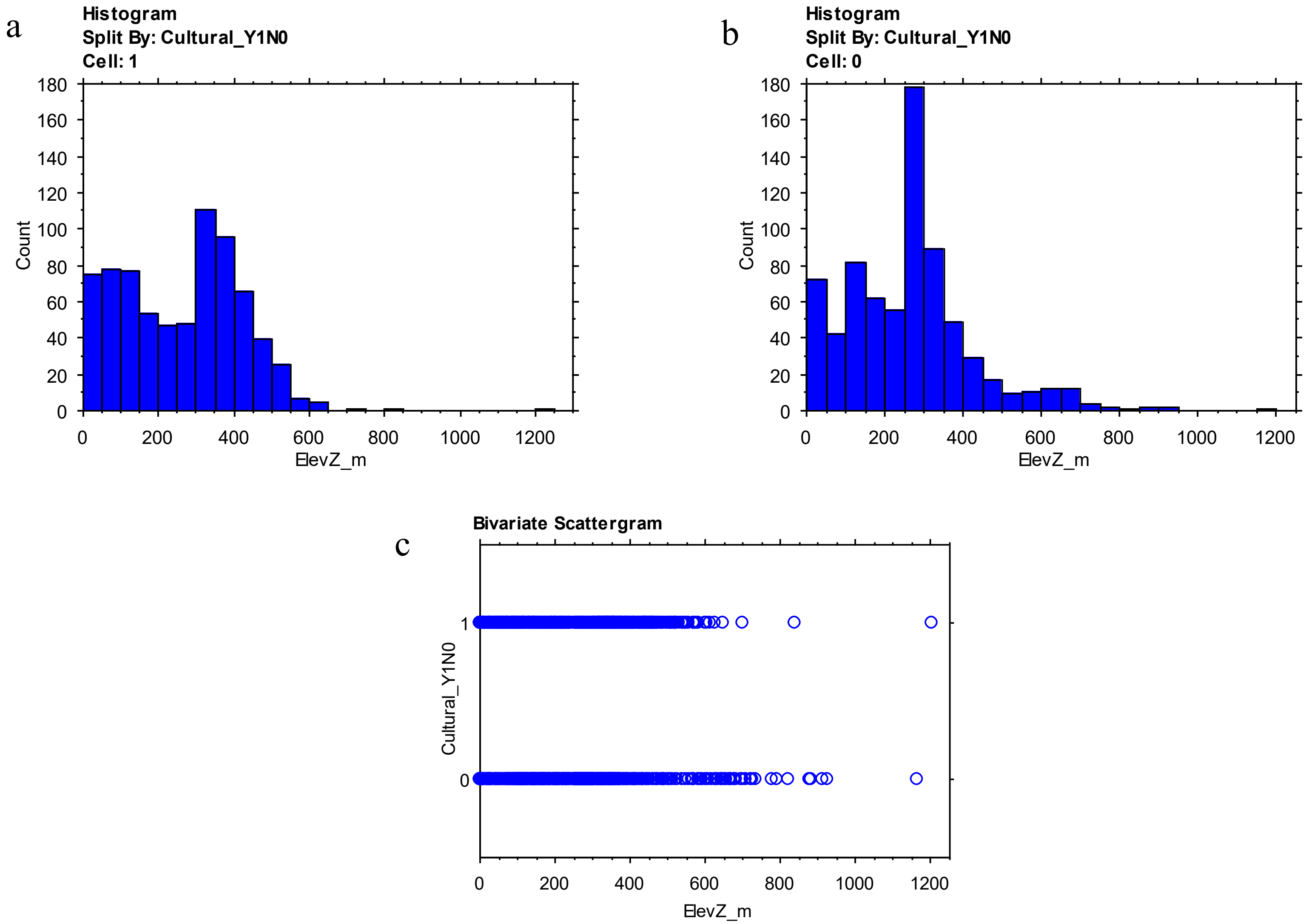


Figure 6. Elevation Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

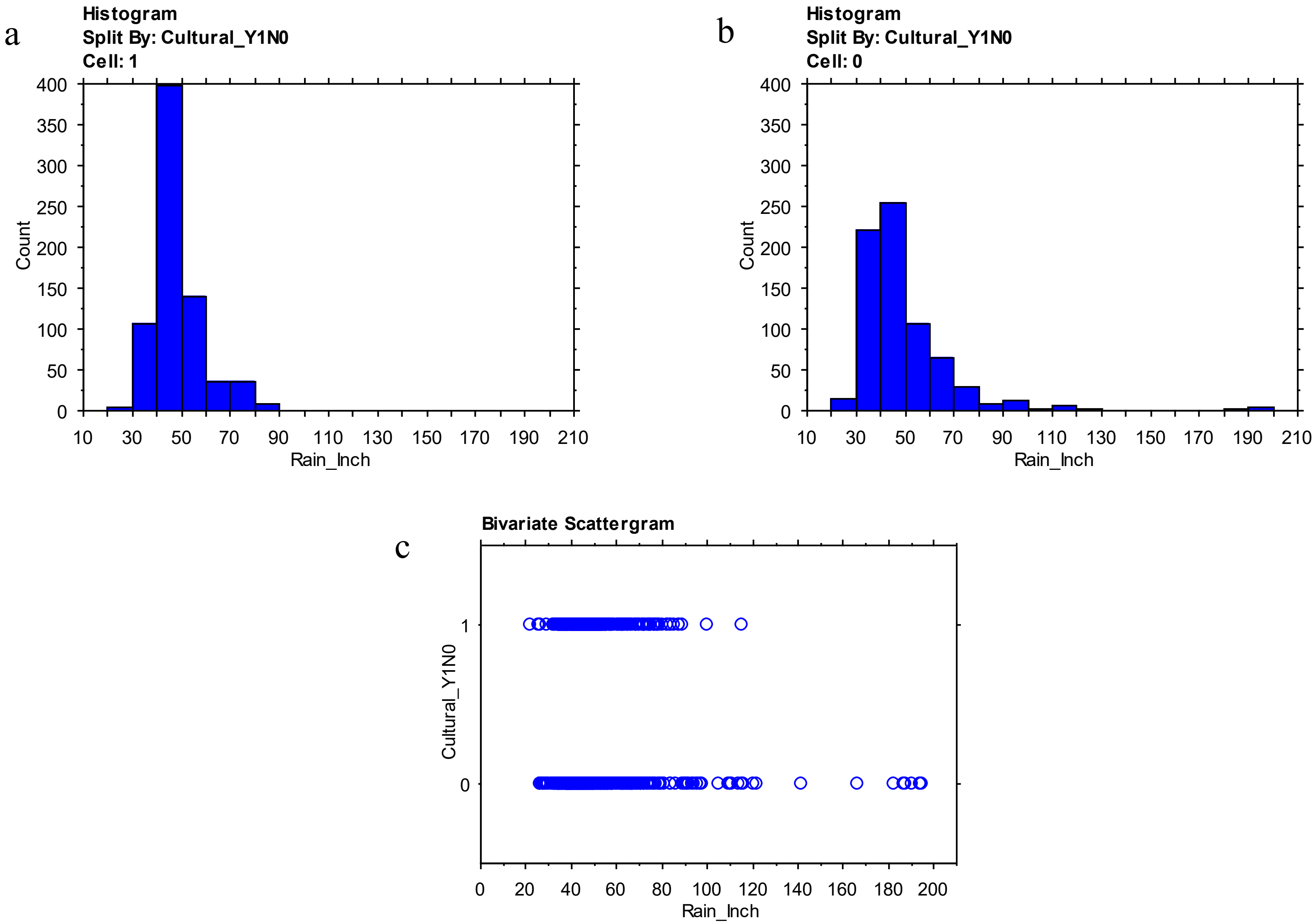


Figure 7. Rainfall Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

between 254 and 107 between 30”-50” per year before decreasing steadily from 99 points until reaching 70” per year. Above 70”, a few non-site points occurred intermittently up to 200” per year. The rainfall distribution for rainfall is a much tighter concentration than the non-sites.

Known site slope occurrence range did not exceed 55° but was visibly more concentrated between 0° and 15° with a dominate peak occurrence between 5-10° slopes (Figure 8). Site occurrence on slopes between 15° and 30° is much lower than the earlier peak and decreased steadily to only a few sites present up to 55°. The non-site point occurrence was most dominate between 0° and 5°, and decreased consistently until terminating 70° slopes.

Site occurrence was dominant at a count of 254 within 50m of a freshwater stream (Figure 9). After 50m the occurrence rate decreases steadily from a count of 113 sites to 600m away at seven sites, then only one or two sites occurred intermittently to 750m and one isolated site near 900m. The non-site occurrence was consistent for the first 200m away at a count near 100 and then decreased steadily from 64 points to trickle out to a single point 1200m away.

For solar radiation categorical zones, site occurrence was highest in two zones between 300-350 WH/m^2 at 333 and between 350-400 WH/m^2 at 213 (Figure 10). The zone between 400-450 WH/m^2 contained 125 sites. The other zones showed marginal representation and no sites fell within the 250-300 WH/m^2 zone. The non-sites were represented between 163 and 209 points for the four zones between 300 to 500 WH/ m^2 while the upper and lower zones had minimal representation.

The majority of the known sites fell within the MLRA 165 soil category for subhumid intermediate mountain slopes at a count of 365 (almost half) of the sample (Figure 11). The next highest occurrence fell into MLRA 167 humid, oxidic soils on low and intermediate rolling mountain slopes at a count of 152, then MLRA 166 very stony land and rock land at a count of 110. The lowest frequency of sites were in the remaining three categories; MLRA 158 semiarid and subhumid low mountain slopes, MLRA 164 humid and very humid steep and very steep mountain slopes, and MLRA 163 alluvial fans and coastal plains occurred at 64, 39, and 9 respectively. The non-site sample was dominated by MLRA 158 at 230, and MLRA 165-167 fell similarly between 175 and 111 points. The remaining two categories fell below a count of 50 points.

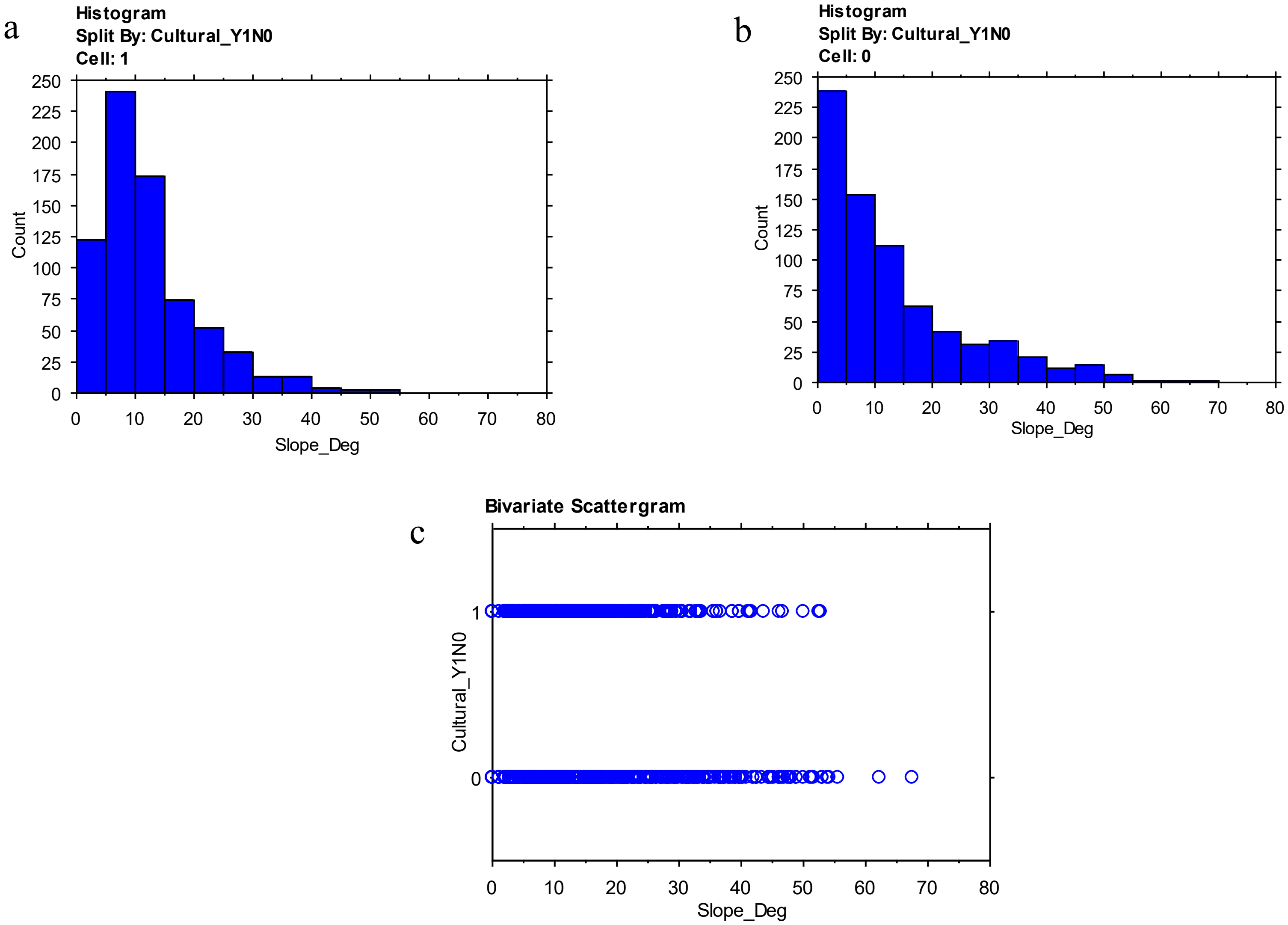


Figure 8. Slope Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

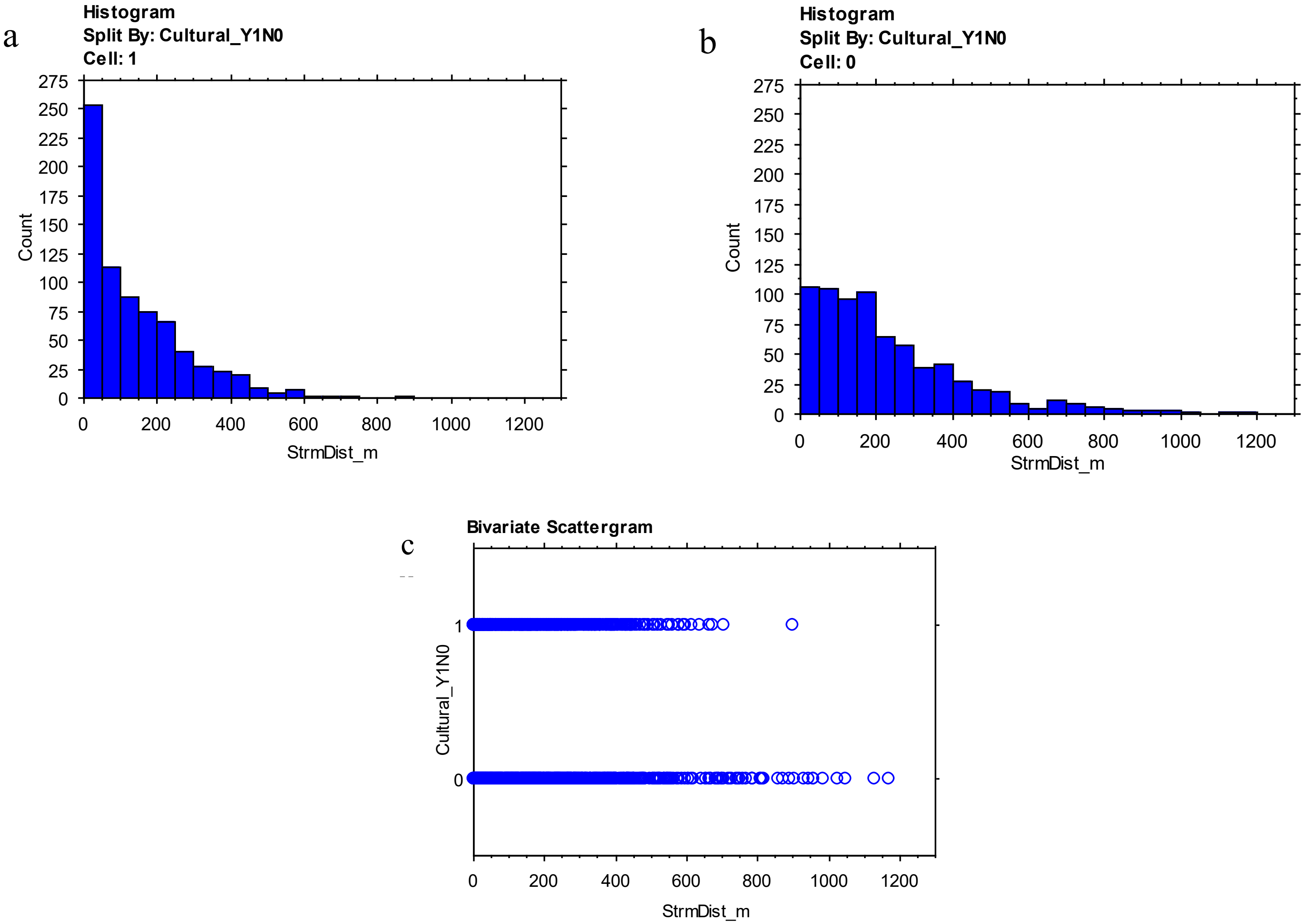


Figure 9. Distance to Stream Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

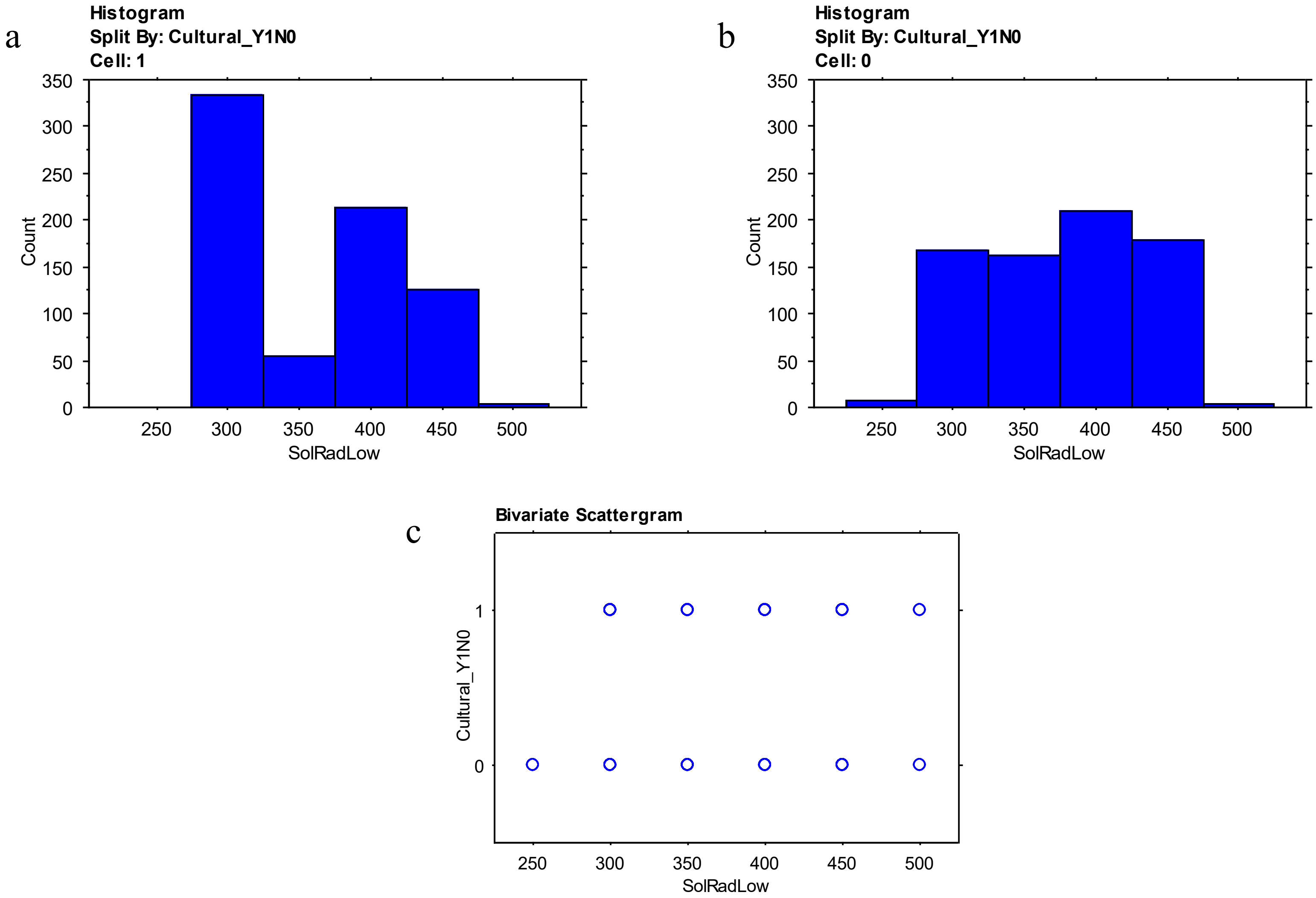


Figure 10. Solar Radiation Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

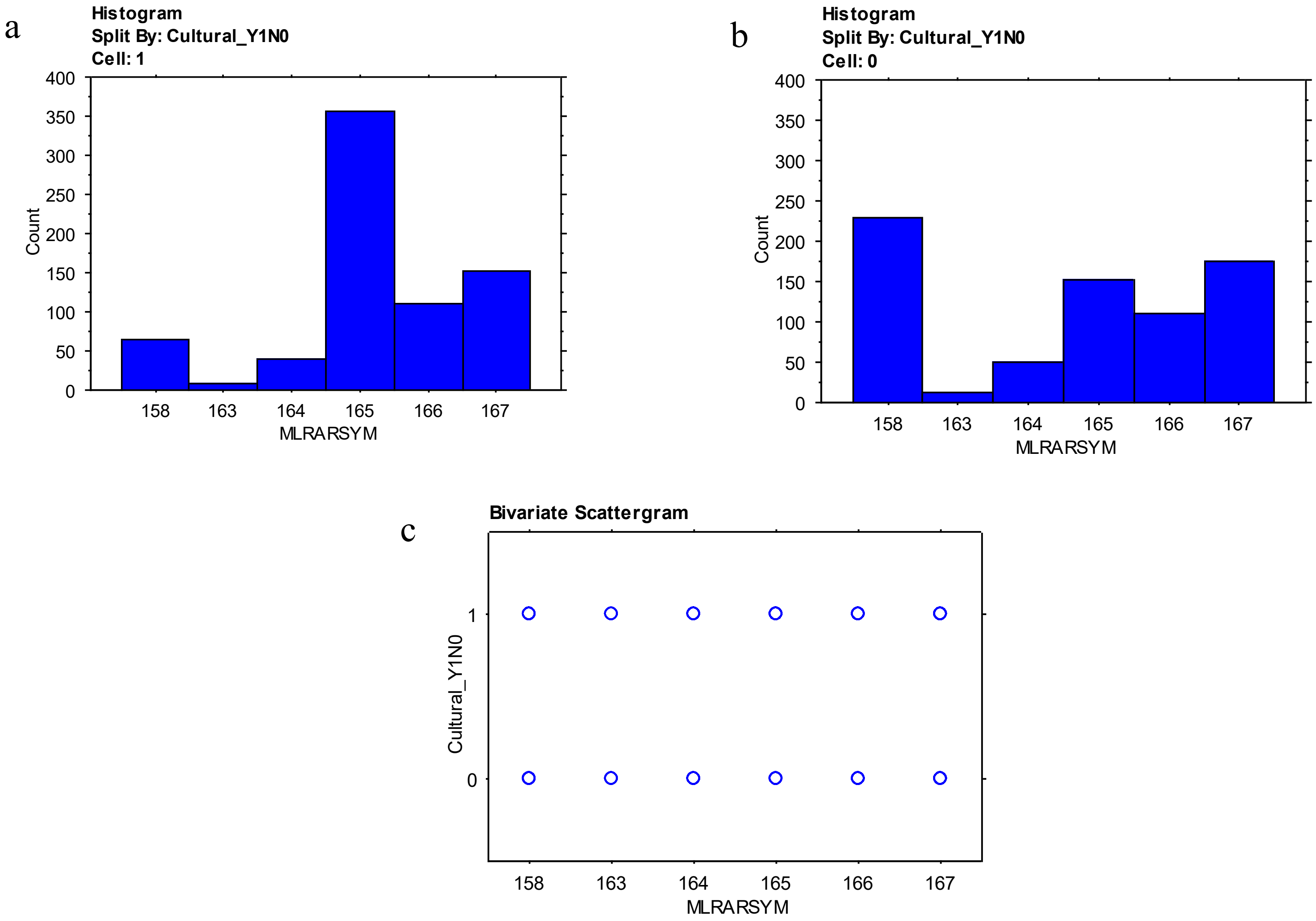


Figure 11. Soil Character Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

For land cover (Figure 12), site occurrence in class 10, Evergreen forest, is clearly dominant at a count of 398 (just over half) of the sample. Class 12, Scrub, is next dominant at a count of 213 followed by class 8, Grassland, at a count of 75. The remaining classes 4, 5 and 6 fell below counts of 23. The non-site sample was also dominated by class 10 at 334 points but had a broader representation of all the other classes, which were all below 88 points.

These figures showed the environmental attribute distributions do not have linear relationships with each other, which indicates each variable is an independent location indicator. The more concentrated distribution of the site points versus the non-sites suggests attribute preferences.

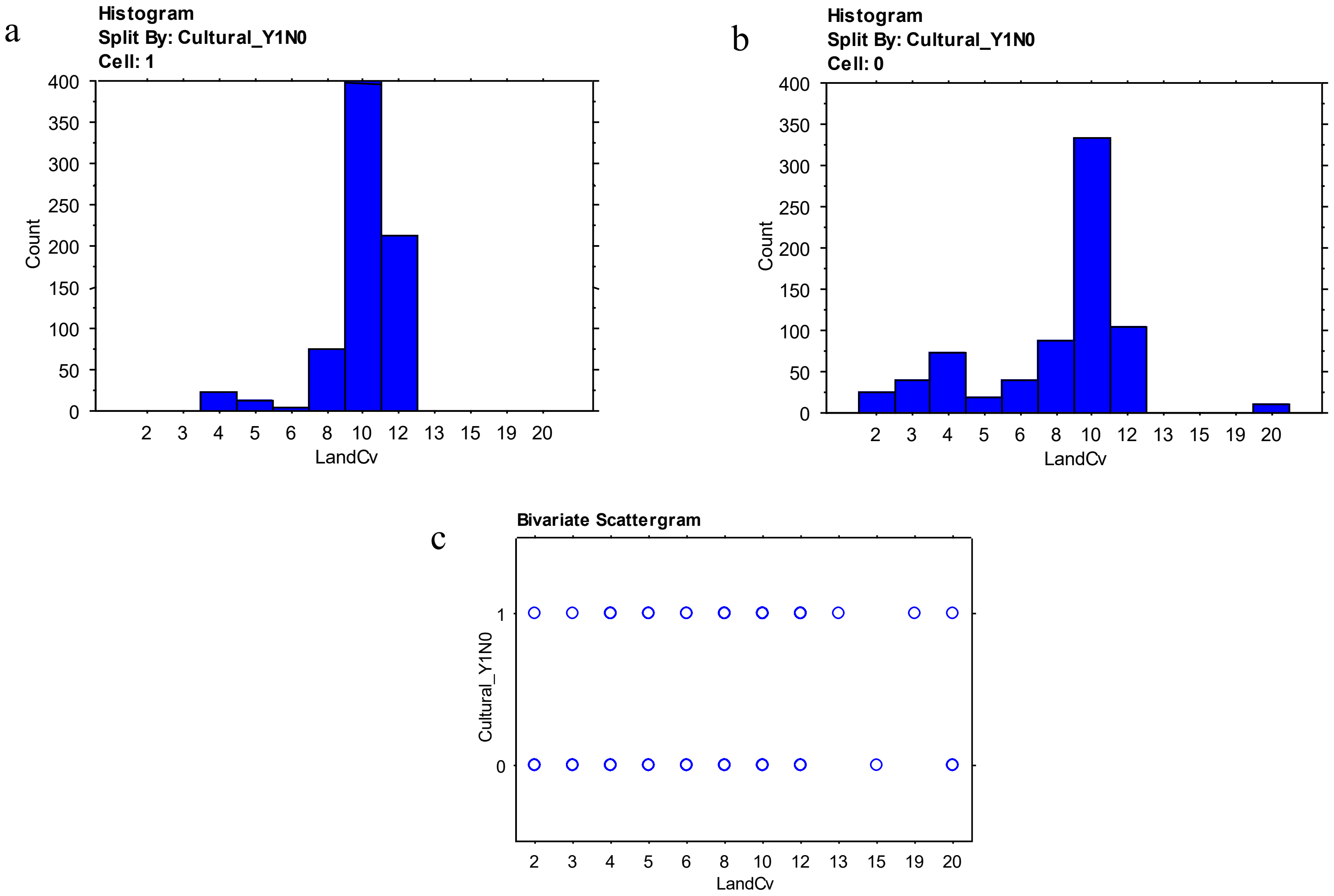


Figure 12. Land Cover Class Comparison Histograms (a and b) and Scattergram (c) for Sites and Non-Sites (N=730 each).

Modeling

Logistic Regression

Logistic regression has proven an effective predictive modeling technique for site locations across a landscape, as indicated in several other site location predictive studies by Parker 1985, Kvamme 2006, Lock et al. 2006, and Mills 2010. By the nature of the logit transformation, it is applicable to non-normal data and does not require a linear relationship (Mills 2010, Moore 2012 and Pierson 2017). Logistic regression is a modeling technique, analogous to linear regression, that examines the relationship between a binary categorical response variable (dependent) and one or more categorical or continuous predictor variables (independent) (Mills 2010 and Moore 2012).

Analysis of the known site data showed that the environmental data was irregularly distributed, did not have linear relationships, and were independent of each other, therefore logistic regression was an appropriate technique for modeling this dataset.

The logistic regression equation analyzes a set of variables and considers the effects of variables in combination with one another. In the logit model, the response variable is log odds (Moore et al. 2012) where the expression P/(1-p) is the odd ratio of the null H_o (that its random point (0) than a site (1)):

$$log\left(\frac{\rho}{1-\rho}\right) = \beta\text{o} + \beta 1x1 + \beta 2x2 + \cdots + \beta 8x8$$

The analysis considers both the known site locations, as well as those locations without sites, to predict where sites are likely and unlikely to occur (Parker 1985). The regression result produces a prediction between 0 and 1, indicating the probability likelihood (as a percentage) of a present resource at a location.

To utilize the logistic regression technique, the random sample discussed previously was combined with the archival sites dataset to create a binomial, or dichotomous, categorical dependent variable using 1 = cultural resource and 0 = random non-site point. The combined training and random sample data (N=1460) were exported into R statistical program in order to fit a machine-learning algorithm. Generalized linear models "glm", are utilized with success in several predictive modeling efforts for ecological studies (Duqu-Lazo et al. 2015 and Huettmann and Diamond 2001) and archaeological studies (Mills 2010). The data were fit to the model using the “glm”, a logistic regression model used to predict a probability by specifying the binomial dependent variable. The model was run specifying the binomial dependent and all eight predictor environmental variables.

The coefficient output for each environmental variable indicates that all eight variables significant and have a strong association with the probability of identifying a site (Table 5) as all attributes fall under the 0.05 p-value (Pr(>|z|)), although shown in scientific notation. In this test, the p-value or Pr(>|z|), is the proportion of the z distribution at that degree of freedom which is greater than the absolute value of the corresponding z statistic. A low p of z-value indicates a strong ability to differentiate between a site and non-site. In the R “glm” binary logistic regression, the log algorithm has a slightly different output, rating slope and elevation as significant unlike the StatView analysis under the paired-means comparison test.

Table 5. Coefficients Output for Environmental Variables in R for N=1460.

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-2.593e+01	3.487e+00	-7.437	1.03e-13
Soils, MLRA	1.929e-01	2.246e-02	8.585	< 2e-16
Solar Radiation, Low	-1.213e-02	1.670e-03	-7.264	3.77e-13
Distance to Stream	-2.681e-03	4.053e-04	-6.615	3.73e-11
Elevation	-3.582e-03	6.394e-04	-5.602	2.12e-08
Land Cover	1.566e-01	2.561e-02	6.114	9.72e-10
Aspect	-1.931e-03	5.396e-04	-3.578	0.000346
Rainfall	-1.699e-02	4.824e-03	-3.522	0.000429
Slope	-1.300e-02	5.934e-03	-2.191	0.028465

Variable Importance was also run in R, “VarImp”, to verify the relative significance of each variable. The scale is fit to the number of variables and correlates to the absolute value of the z-value, with the highest number being more important. Table 5 reflects the order of variables by importance.

Performance Assessment

To assess the goodness of fit, an analysis of variation (ANOVA) test (likelihood ratio chi-squared) (Huettmann and Diamond 2001) to analyze the table of deviance is a common option.

This test compares the likelihood of the data of the full model against that of fewer predictors. It is expected that removing predictors will result in a model with a lower likelihood, but it is useful to test that the difference in the model fit is statistically significant and a good model fit. For this test in R, the Pr (>Chi) is the p-value, of which all fall below 0.05 and interpreted to have a significant effect on the model. The NULL indicates the response predicted by the model with nothing but the intercept. The residual deviation indicates the response predicted by the model adding independent variables, which in this case was first to last in Table 6. The deviance decrease indicates the effect of adding each independent variable. The residual deviance decreases consistently with each addition, however, less so with the last three variables.

Table 6. Table of Deviance.

Variable	**Degree of Freedom (Df)**	**Deviance**	**Residual Df**	**Residual Deviation**	**Pr(>Chi)**
NULL			1459	2024.0	
Soils, MLRA	1	88.5	1458	1935.5	<2.2e-16
Solar Radiation, Low	1	59.8	1457	1875.7	1.026e-14
Distance to Stream	1	48.0	1456	1827.7	4.190e-12
Elevation	1	77.5	1455	1750.2	<2.2e-16
Land Cover	1	42.3	1454	1707.9	7.791e-11
Aspect	1	14.7	1453	1693.2	0.0001275
Rainfall	1	16.4	1452	1676.8	5.180e-05
Slope		4.9	1451	1672.0	0.0273654

While data portioning can reduce the size of the training data and result in over-estimates of the actual error rates, averaging of K-fold partioning can make the model accuracy estimate less dependent on a single partition (Fielding and Bell 1997). The model was tested by running five K-fold cross-validation iterations to assess how well the model performs in predicting sites and non-sites. This included scrambling the binary dependent column of 1/0 to ensure an equal distribution occurred in each training and test sample. Five iterations were used, rather than the common 10-fold, because the sample size was sufficiently large. Each K-fold iteration used a training sample of 1168 and test sample of 292. The K-fold outputs are summarized in Table 7. In two iterations, slope was not significant. The average of the K-fold accuracy test resulted in

70% probability of predicting the presence of the sites and non-sites from the N=292 test sample of each iteration.

Another common performance measurement for a binary classifier is to plot the receiving operating characteristic (ROC) curve that compares the true positive rate against the false positive rate. The area under the curve (AUC) is calculated, for which a good predictive ability will have a value closer to one than 0.5 (Fielding and Bell 1997, Roberts et al. 2010 and Duqu-Lazo et al. 2015). The average of the five cross-validations on the test samples resulted in a 76% AUC (Table 7 and Figure 13). These results indicate this model would be useful for predicting sites or non-sites.

Table 7. Summary of K-Fold Cross Validation Iterations.

Cross-Validation Iterations (N=292)	**Coefficients & ANOVA Summaries**	**Accuracy**	**AUC**
		Stronger *closer to 1*	*Stronger* *closer to 1*
Overall (N=1640)	Coefficient: Intercept + 8/8 variables Pr(>|z|) Significant ANOVA: 8/8 variables Pr(>Chi) Significant	NA	NA
K1: 1169-1460 test	Coefficient: Intercept + 8/8 variables Pr(>|z|) Significant ANOVA: 8/8 variables Pr(>Chi) Significant	0.71	0.75
K2: 1-292 test	Coefficient: Intercept + 7/8 variables Pr(>|z|) Significant (not Slope) ANOVA: 7/8 variables Pr(>Chi) Significant (not Slope)	0.70	0.76
K3: 293-584 test	Coefficient: Intercept + 8/8 variables Pr(>|z|) Significant ANOVA: 8/8 variables Pr(>Chi) Significant	0.67	0.72
K4: 585-876 test	Coefficient: Intercept + 8/8 variables Pr(>|z|) Significant ANOVA: 8/8 variables Pr(>Chi) Significant	0.72	0.80
K5: 877-1168 test	Coefficient: Intercept + 7/8 variables Pr(>|z|) Significant (Not Slope) ANOVA: 7/8 variables Pr(>Chi) Significant (not Slope)	0.70	0.76
Averages		0.70	0.76

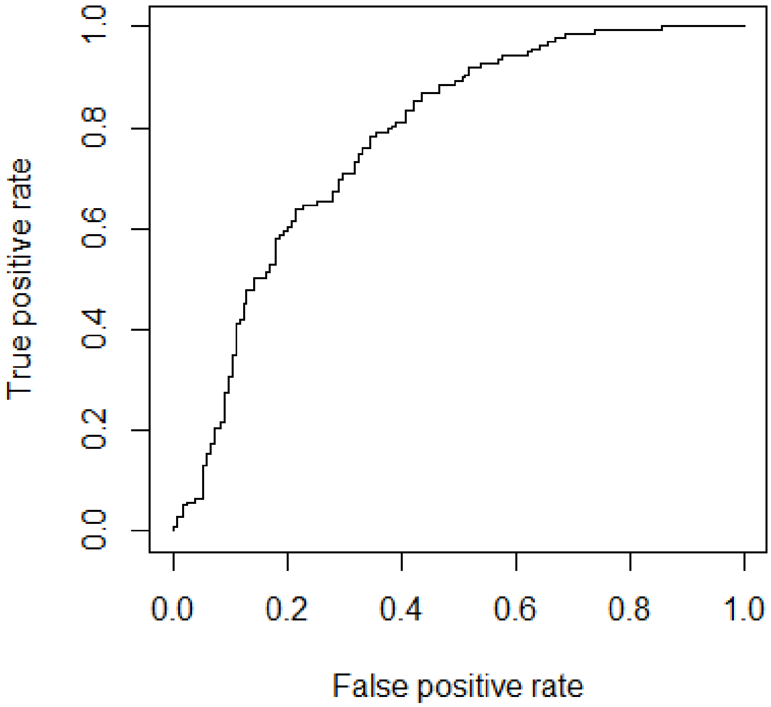


Figure 13. Plot of the ROC Curve and AUC of 0.76.

Model Output

A final step to build an output raster with probability rated cells was conducted in MATLAB. Once all statistical tests confirmed the data were valid and diagnostic tests confirmed the fit of the "glm" regression model for useful probability estimates using StatView and R, the data were imported into MATLAB to create a probability raster. The same spreadsheet for the combined sites and non-sites (N=1460) was loaded. The regression model was replicated using the five attributes indicated as significant contributors; aspect, rainfall, elevation, distance to stream, and solar radiation. The categorical variables soil and land cover were omitted since they are coded and not compatible with the raster output. Additionally, slope was omitted after an initial run in MATLAB during calibration of the model, which yielded a p-value of 0.62 and therefore determined not a significant contributor to the model. This was not unanticipated as slope was not significant in several of the K-fold coefficient tests. The matching rasters, as described previously in Data Processing, were also imported into MATLAB as predictors. The predicted output from the regression model was calculated and plotted for each cell based on the spreadsheet data and rasters.

The resulting 10m geotiff based on the remaining five attributes was exported as a geotiff. Each raster cell was assigned a value between zero (0), no likelihood, and one (1), high likelihood, representing the percentage of likelihood to predict the presence of a site. The map was coded to WGS84 Zone 4 North spatial reference for the Island of Oʻahu.

Figure 14 shows the final likelihood output map of Oʻahu for grid cells rated 0 to 1 (or 0-100%) likelihood to contain a site. A general review of the output appears to be strongly influenced by the distance to stream which very consistently shows higher likelihoods (lighter color) nearest the streams. The northeast facing orientation, especially on the coast, appears a higher likelihood apart from a few locations. High elevations appear to support the lowest likelihood areas along the ridges, however, there are several southwest coastal zones indicated as low. This contradicts previous research observations that the coastal areas were preferred. This may be a result of limited coastal data to train the model.

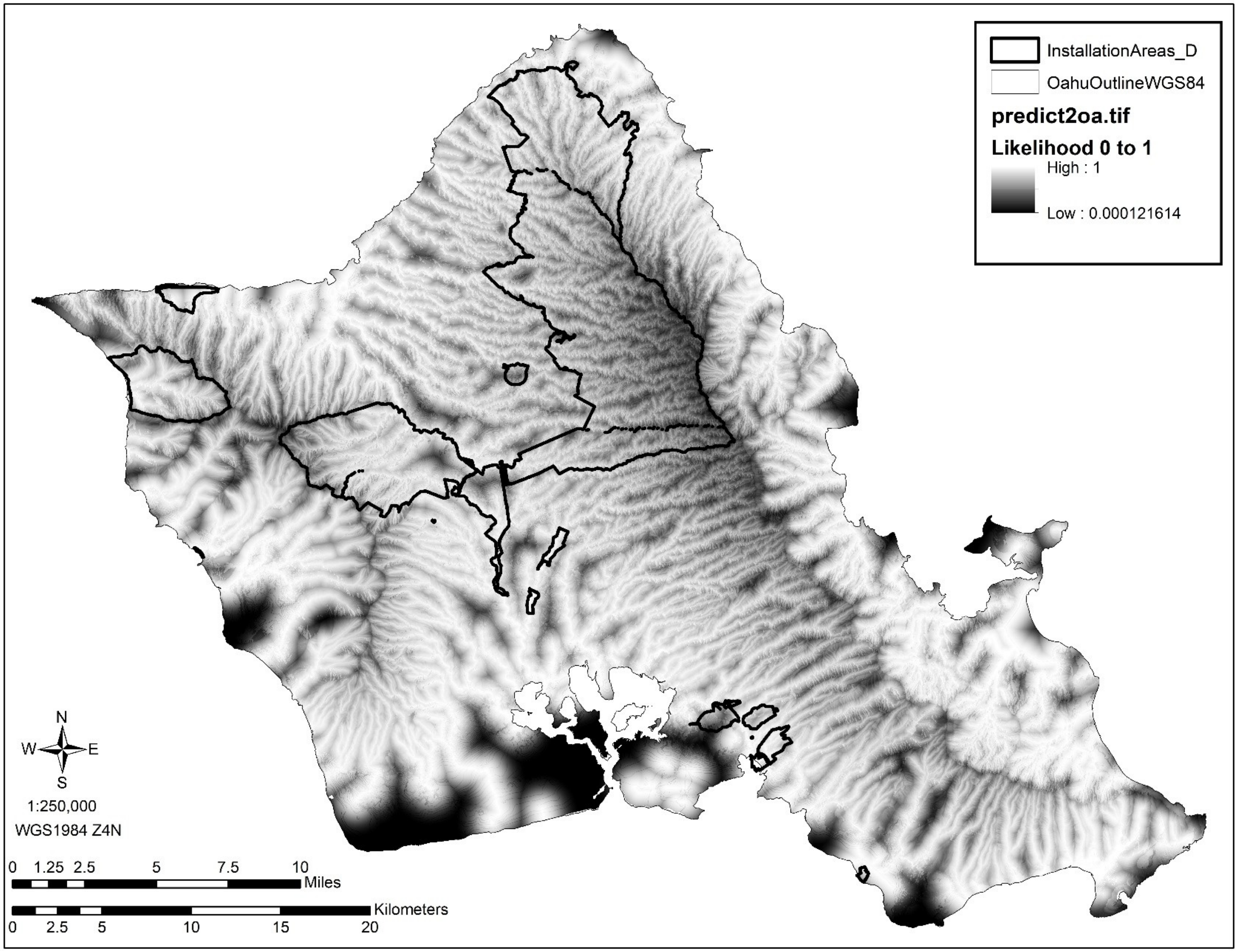


Figure 14. Predicted Likelihood Map for O‘ahu Island.

Model Validation

The reserved known test site sample (N=89) from the three surveys and random non-site sample (N-89), were validated against the resulting probability map defined by the training sample. The likelihood probability values were extracted to the point feature classes for both sites and random/non-sites by a similar process as described earlier but using Extract Multi Values to Points, assigning the cell center value to the point.

The descriptive statistics indicated the central tendencies were nearly centered (Table 8). The standard deviation was less than both means; however, the mean for the sites is slightly smaller, but unlikely enough to qualify a difference in variability. Of the 89 survey test points, the points fell in likelihood grid cells between 0.68 to 0.94 (68-94%) and the random non-sites fell between 0.58 to 0.93 (58-93%). The mean for the test sites is slightly higher with a slightly tighter distribution range than the random sample. For reference, the distribution range within the survey areas is between 0.55 to 0.95 (55-95%). The distribution range of the probability across the island is between 0.0001 to 0.9998 (0-100%).

Table 8. Descriptive Statistics for Validation Sample (N=89).

Site Class	Mean	Median	Mode	Std. Dev	Range	Min	Max
Sites	0.84	0.85	None	0.06	0.26	0.68	0.94
Non-Sites/Random	0.82	0.83	None	0.08	0.35	0.58	0.93

The distribution of sites and random/non-sites are displayed in histograms for comparison (Figure 15) and indicates a greater range spread for the random non-sites (0) than the sites (1). Between 0.82 and 0.90 shows a peak occurrence of sites representing 59.6% of the sites in the sample. The point occurrence of the site sample is much more consistent and clustered than the random non-site sample viewed in 2% increments.

This random sample was not used for any previous analysis or examined for overlaps with known sites. The test sites and random non-sites were overlaid on the output probability geotiff (Figure 16) in ArcGIS. For simplified display, the likelihood probabilities were broken into

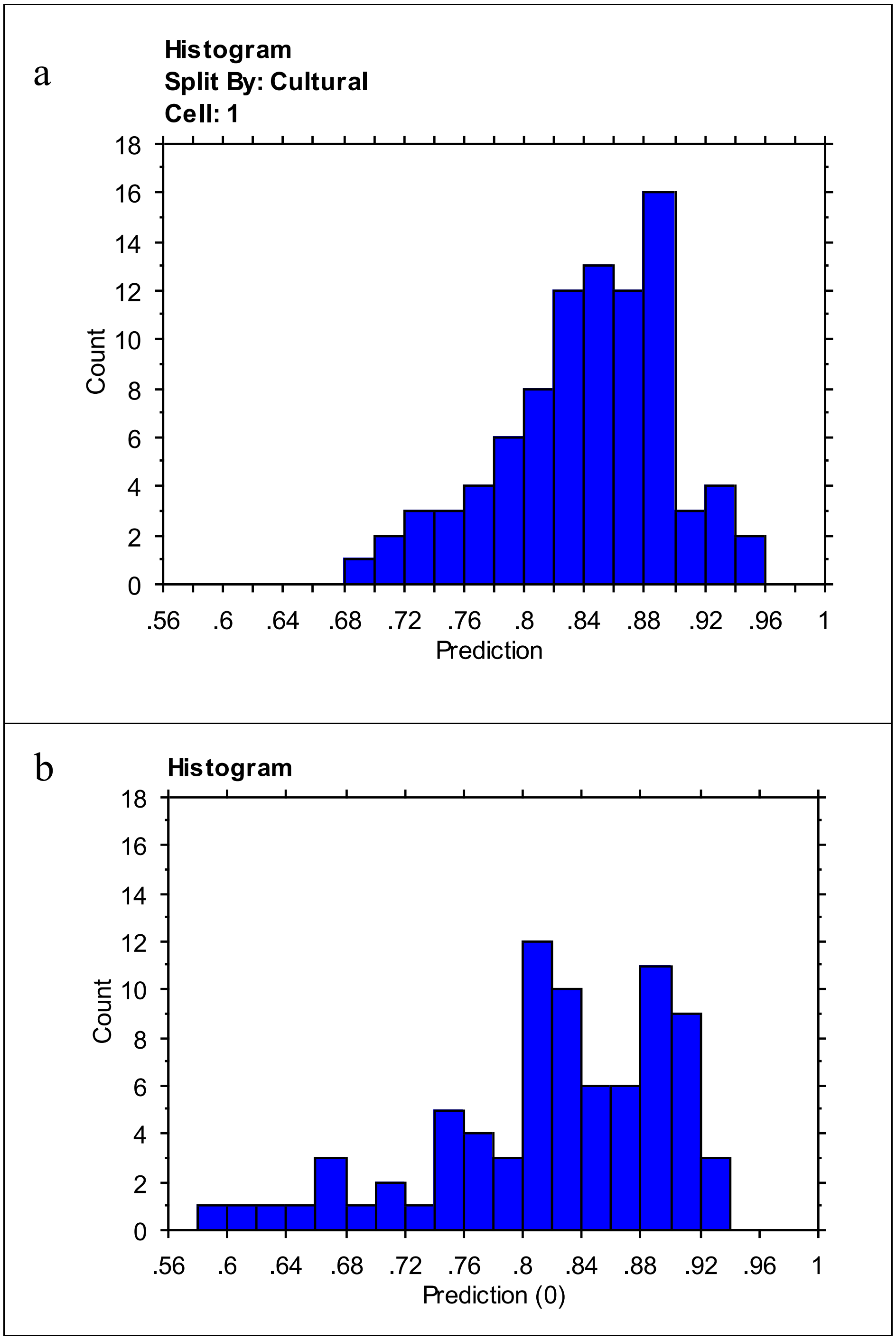


Figure 15. Distribution Comparison of Test Validation Points, Sites (a, N=89) and Non-Sites (b. N=80).

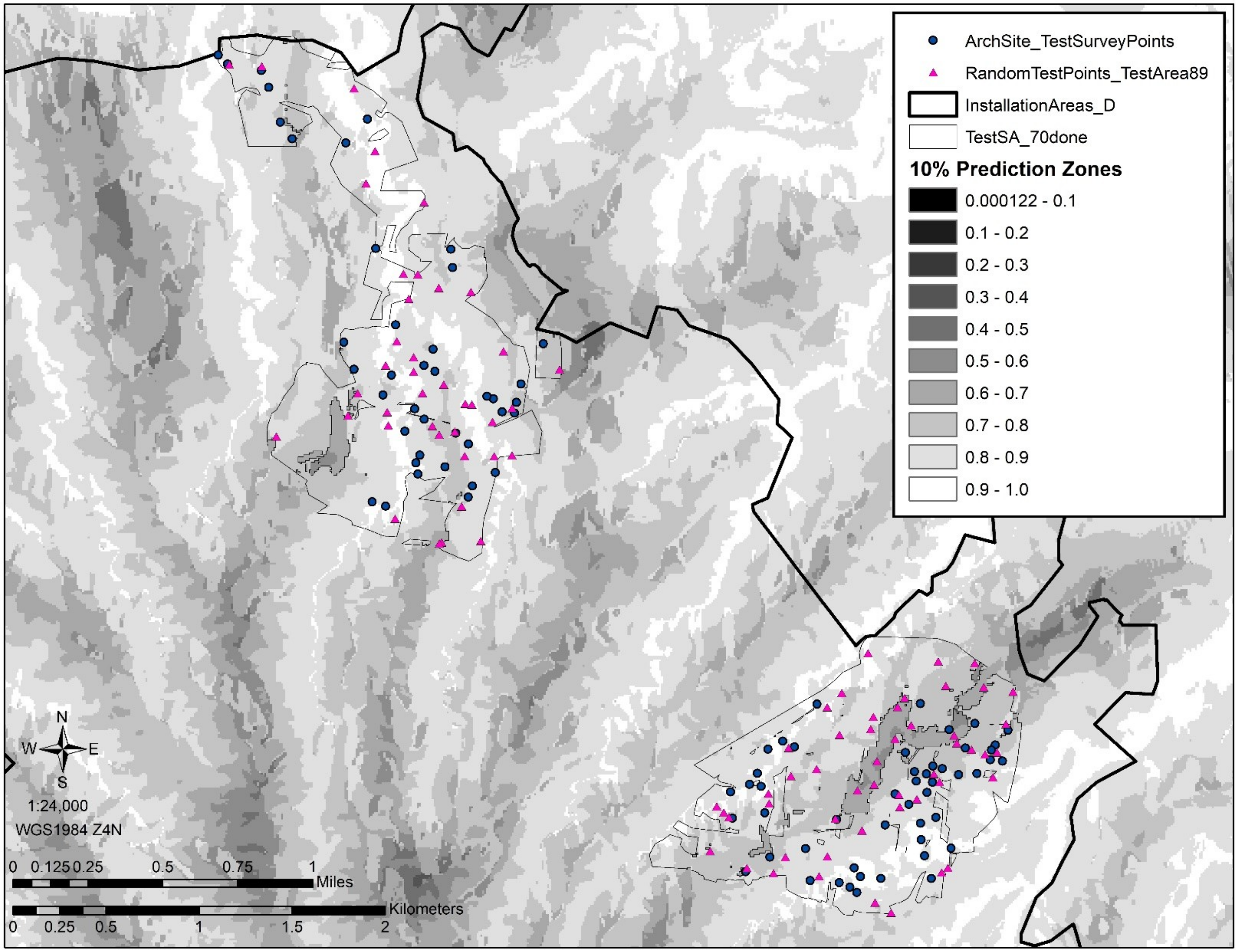


Figure 16. Survey Results Test Sample Sites and Non-Sites Overlaid with Probability Zones.

10% increments. This spatial overlay shows nine random points fall within trusted site polygons but no closer than 10m (a cell size) from a test site point, and another five points fall within 30m of a known site point. The cell values for all these sites rated high between 0.76 to 0.91. There are also several sites surrounded by site points and trusted polygons, but not within. A sample of these points fell between 0.66 and .84, somewhat lower than known sites. The nine correlating with sites were removed from the sample to truly represent non-sites. The decreases were mostly in the point occurrence peak, resulting in an almost bimodal distribution (Figure 15 b), however did not alter the descriptive statistics shown in Table 8.

Given the limited range of environmental variability and limited range of likelihood distribution within the test survey polygons, the ability to confirm the reliability of the model is challenging. The histograms show much overlap in point distribution range, but the sites did fall into higher likelihood cells than the non-sites. The tighter histogram range spread for sites suggests some model reliability.

CHAPTER 5: CONCLUSION

This research explored eight environmental attributes with potential to provide insight on desired areas chosen in the past for human activity. Aspect, elevation, rainfall, slope, distance to fresh water, solar radiation, soil character, and land cover class were each analyzed for their association and relationship to known sites present on USAG-HI lands on the island of Oʻahu. Outcome of the research indicates that five attributes: aspect, elevation, rainfall, distance to fresh water and solar radiation, all have significant associations with site location.

Based on variable importance test conducted in R, which reportedly accommodates categorical data input, soil class was the most important variable. Solar radiation was second most important, followed by distance to stream, elevation, land cover, aspect, rainfall, and finally slope. Slope was found to have the lowest significance value in R, and was found not significant in MATLAB.

The MATLAB analysis was run without the categorical attributes of soils and land cover after consideration of the data coding, despite a high importance rank in R. Slope was also omitted because of its lack of significance. Distance to stream was the highest ranked discrete variable attribute and the final output map from MATLAB visually shows obvious incorporation of stream data. Modeling the five significant attributes in MATLAB resulted in a locational probability model with a 10m output raster of rated grid locations across the island of Oʻahu between 0 -100% likelihood of a site/human activity.

In seeking the goal to define a reliable probability model for USAG-HI lands, the data was examined for natural breaks. The peak site occurrence of the validation sample captures 59.6% of the sites between 0.82 and 0.90 likelihood. It is assumed the probability likelihood cells up to 1.0 would also predict sites if future surveys were in the range, which extends the likelihood occurrence to 69.7% of the validation sample. If the average 0.76 AUC from the K-fold model assessment diagnostic is an accurate estimate of the predictive ability, which would again expand the likelihood occurrence to include 89.9% of the validation sample. This 0.76 almost follows the middle histogram distribution natural break, which falls at 0.78. There is also justification, considering the limited coverage of the test survey area, to include the full spectrum of site occurrences down to 0.68 as a lower threshold of expectation.

For the purposes of USAG-HI land management and planning, the data was refined to installation areas. Of the 59,652.8 acres within installations, 45,434.9 acres remain to be surveyed. Of the higher likelihood probability areas between 0.82 to 1.0, 12,881.5 acres (28.5%) are unsurveyed. If lower likelihood probability areas are considered, 0.76 to 0.82 adds another 8,404.8 acres (18.5%) and 0.68 to 0.76 adds yet another 9,384.9 acres (15.7%) to need survey. This leaves 26,369.9 acres (32.3%) below 0.68 likelihood. Figure 18 shows the likelihood probability zones above .68 and previously surveyed areas within the installations. Figure 18 focuses on the northern installations with more area to be surveyed. The southern installations only require small areas surveyed.

Based on these results, this type of modeling could guide expectations in research designs and early development planning to anticipate the need for advance survey and support funding requests as desired. The high likelihood probability zones could help anticipate the duration of survey needed if it is all within high verses low probability. This model will not eliminate surveys, but could provide priority areas to focus survey on in preparation of development. Survey is still necessary to document and evaluate the sites.

The use of GIS and remote sensing technology was critical for the development of this model. GIS was used to process the bulk of the data compilation and manipulation between data types such as spreadsheets, rasters and feature classes, and data display. The environmental attributes were fairly easy to manipulate and categorize into useful units for comparison and display. Each attribute required its own analysis and breakdown.

Many of the imported raster or feature class data were derived from remotely sensed technologies, and professionally processed into easily, or fairly easily, accessible products. Processing these products to investigate association with known site points provided insight on the relationship of the environmental attributes. The rank of each attribute provides a start for exploring the connections between them and sites and combinations of them. Because of the large known training sample size, the environmental attributes appear to be reliably predictable, however the model requires field validation in more diverse settings and challenged in areas not defined in the high probability zones.

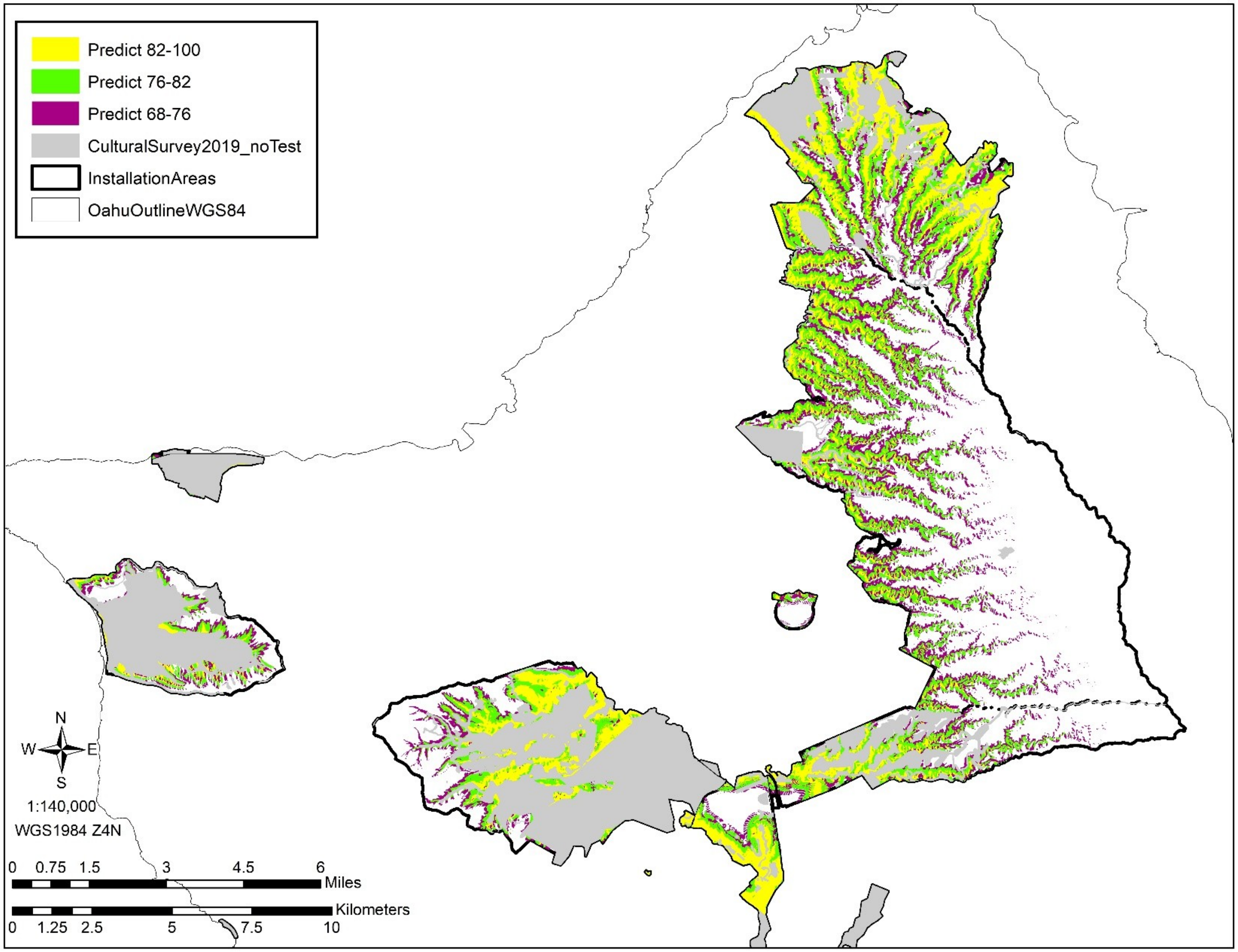


Figure 17. USAG-HI Installations Showing Unsurveyed Areas Within the .68 – 1.0 Probability Zone.

While GIS is a critical tool for data processing and manipulation, the final modeling would not have been successful without utilizing common outside programs that provided more powerful functionality. The statistical options in ArcGIS were not found to be as powerful or capable of handling the various combination of data analysis. The ability of MATLAB to perform the statistics, process images, and convert them into graphical output reduced a significant workload if it were attempted in ArcGIS.

This type of modeling is very cost effective, however somewhat time consuming. More refined data may be incorporated if available, but is not necessary to create a model to test. This model is based on O‘ahu Island, but the process could be applied for other areas based on the available site data. The same processes can be applied to the dataset to tailor the probability zones.

CHAPTER 6: DISCUSSION AND FUTURE DIRECTIONS

In the process of this research, several shortcomings pertaining to data were identified and additional resources came to light that may increase the utility and functionality of such a model and are briefly reviewed.

Dataset Limitations

While the USAG-HI dataset was large, it is limited in general coverage. Very few surveys have occurred in the higher uplands to provide insight to a probability model. The USAG-HI dataset had very minimal coastal representation just by the nature of its land-holdings. As mentioned in the earlier literature review, coastal sites are extremely common for early settlement and it is a limitation of the available dataset that very little of the property owned by the Army falls within this zone. Other work conducted on the island beyond the army installations, have proven that many sites are present in and along the coastal regions.
A visual comparison with a very outdated and known poor quality-controlled dataset of the island shows that there are sites in these areas (or at least the vicinity). This suggests this model requires some adjustments that could improve the true predictive function. However, since there are few coastal lands the USAG-HI manages, the predictive utility may be suitably accurate for the uplands.

A general overview of the source installations for the dataset shows an irregularity of site density that might affect the model. The majority of the sites occur in KTA and Dillingham Military Reservation (which are generally oriented north) and Schofield (which is generally oriented east). Several very intensive surveys have occurred in these installations yielding a great many sites compared to other areas, but these installations have the largest tracts of training lands requiring such attention. Installations MMR, Pililaau Army Recreation Center (PARC), Fort Shafter and Fort DeRussy are all generally oriented west which, has the least dominant direction in the aspect dataset but also has the least site representation and less acreage. PARC and Fort DeRussy both have very few, but large sites.

Another look at the known site data could improve the model. This research focused on only site point data, generally a centroid of the site. Most sites are larger than a 10m by 10m grid cell. Initial attempts to use polygon data were unsuccessful and the current USAG-HI is still refining accurate site boundary polygons. Using polygons, or perhaps creating additional site points per cell within each site boundary to accommodate the site size, would increase the point sample size and provide useful environmental variation for each site. In this research for example, only one centroid point was used to represent a 50m by 20m site. If a point were created at each cell center within the site polygon, the site would be represented by 10 points, which would offer more information about the site environment.

Creation of the randomly generated sites could also be refined. In this research, the random points were created within the survey areas, however that included all the known sites. Future analysis would benefit from subtracting the site polygons from the survey areas to ensure there no random non-site points created within sites terrain. This would likely provide greater differences between the site and non-site environmental attribute data.

A reevaluation of the categorical datasets might discover values that can be ranked or other discrete number data that would offer more flexibility to contribute significantly a model. In the case of land cover, more refined, or differently expressed, data may contribute more to the model. What was used was general vegetation canopy type, but it did not indicate primary, secondary, or invasive species coverage. A primary forest seems likely to support more intact cultural resources.

Additional Environmental Attributes

The model itself could benefit from additional environmental variables. The land cover class indicates built environment but not specifically disturbed areas that would have a low likelihood of containing cultural resources. O‘ahu is very developed and a review of a series of aerial photographs and other remotely sensed imagery could eliminate some areas of higher probability if construction, intensive agriculture or development is observed, since the islands' cultural deposits are relatively shallow.

Distance to marine resources, such as the coast in general, or specifically, distance to resource-rich harbors, may also prove significant contributors.

Addition of a soil fertility classification could shed light on identification of agricultural verses non-agricultural sites.

Utilization of Other Imagery

LiDAR was explored for this research, but what was available was found to be too low of resolution to provide a DTM better than the NED DEM or that could show surface relief to the extent that constructed surface features could be observed. Very recently, a new effort is underway to obtain Q1 level LiDAR over training the ranges.

Additional Model Validation

As the three test sample survey areas were defined based on training needs, and not as a model test, the areas were found not to be a good test of the model. A test area, or several areas, should be planned to target various areas of the probability to really test the likelihood prediction. Inclusion of subsurface testing in the high and low probability areas, and areas previously surveyed (that yielded no site surface indicators), would also challenge the model. Unfortunately, the currently funded survey is already slated for an area that is also largely within the 0.68 and greater likelihood probability area only.

Further Directions

Initial attempts at modeling looked at the frequencies of the known sites for each site type; agricultural, habitation, combination, undetermined, general historic and military. The frequencies indicated that there was some variability between types; and with a proven machine-learning model, identification of types of sites would be useful for all the same reasons as this model. Additionally, there are many sites currently with undetermined typologies. If such a model were possible, many sites could be revisited with better research questions to shed insight on appropriate typologies based on their environmental attributes.

REFERENCES CITED

Anderson, Lisa (1998) *Cultural Resources Management Plan Report O‘ahu Training Ranges, Island of O‘ahu, Hawaii*. Prepared for US Army Engineer District, Honolulu, Contract DACA83-95-D-0006, Delivery Order No. 0001, Prepared by Ogden Environmental and Energy Services Co. Inc., Honolulu, Hawai’i, August 1998).

Anderson, Lisa and S. Williams (1998) *Historic Preservation for the Kahuku Training Area, O‘ahu Hawaii*. Prepared for US Army Engineer District, Honolulu, Contract DACA83-91-D-0025, Delivery Order No. 0017, Prepared by Ogden Environmental and Energy Services Co. Inc., Honolulu, Hawai‘i, April 1998).

Clover, T. J. (1995) *Pocket Ref.* Sequoia Publishing. ISBN 978-1885071002.

Connolly, James, and Mark Lake (2006) *Geographical Information Systems in Archaeology*, Cambridge University Press, New York.

Cordy, Ross (2002) The *Rise and Fall of the O‘ahu Kingdom, A Brief Overview of O‘ahu’s History*. Mutual Publishing, Honolulu, HI.

Dalla Bona, Luke (2000) *Protecting Cultural Resources through Forest Management Planning in Ontario Using Archaeological Predictive Modeling*, Chapter 5 of Practical Applications of GIS For Archaeologists, Edited by Westcott and Brandon.

Dega, Michael F., and L. McGerty (2002) *A Cultural Resources Inventory Survey, Phase II, of the U.S. Army Kawailoa Training Area (KLOA)*, for the U.S. Garrison, Hawai‘i, Ecosystem Management Program, O‘ahu, Island, Hawai‘i, Traditional and Historic Setttlement of the Kawailoa Uplands. Prepared for US Army Corps of Engineers, Honolulu District, Fort Shafter, Contract DACA83-95-D-0004, Task Order No. 0022, Prepared by Scientific Consultant Services / Cultural Resources Management Services, Honolulu, Hawai´i, Revised May 2002.

Desilets, Michael D., Maria Kaimipono Orr, Christophe Descantes, Windy McElroy, Amanda Sims, Dana Gaskell, and Marion Maiko Ano (2011) DRAFT- *Traditional Hawaiian Occupation and Lō Ali´i Social Organization on O´ahu’s Central Plateau: An Ethno-Historical Study*, Prepared for US Army Engineer District, Honolulu, Contract W9128A-05-D-0007, Task Order No. 0004, Prepared by Garcia and Associates, Kailua, Hawai´i, January 2011.

Dixon, Boyd, Dennis Gosser, Scott Williams, Jennifer Robins, Constance O'Hare, Laura Gilda, and Stephan Clark (2004) *Final Report, Cultural Resources Survey of Selected Lands Naval Magazine Pearl Harbor, Lualualei Branch, Island of O`ahu, Hawai´i*. Prepared for the Department of the Navy Pacific Division, Naval Facilities Engineering, Pearl Harbor. Contract N26742-97-D-3502 Task Order 0011.

Doneus, M., and C. Briese (2006) *Digital terrain modeling for archaeological interpretation within forested areas using full-waveform laserscanning*, The 7th International Symposium on Virtual Reality, Archaeology, and Cultural Heritage (VAST).

Duncan, Richard B., and Kristen A. Beckman (2000) *The Application of GIS Predictive Site Location Models within Pennsylvania and West Virginia*, Chapter 3 of Practical Applications of GIS For Archaeologists, Edited by Konnie L. Westcott and R. Joe Brandon.

Duque-Lazo, J., H. van Gils, T.A. Groen, and R.M. Navarro-Cerrillo (2016) *Transferability of species distribution models: The case of Phytophthora cinnamomic in Southwest Spain and Southwest Australia*, Ecological Modeling, Volume 320, pgs 62-70.

Ford, Anabel, Keith C. Clarke, Gary Raines (2009) *Modeling Settlement Patterns of the Late Classic Maya Civilization with Bayesian Methods and Geographic Information Systems*, Annals of the Association of American Geographers, Volume 99(3) 2009, pp. 496-520.

Fielding, Alan H., and John F. Bell (1997) *A review of methods for the assessment of prediction errors in conservation presence/absence models*, Environmental Conservation, Volume 24 (1): 38-49.

Giambelluca, T.W., Q. Chen, A.G. Frazier, J.P. Price, Y.-L. Chen, P.-S. Chu, J.K. Eischeid, and D.M. Delparte, 2013: *Online Rainfall Atlas of Hawai'i. Bull. Amer. Meteor. Soc.* 94, 313-316, doi: 10.1175/BAMS-D-11-00228.1.

Green, Roger (1980) Māhaha Before 1880 A.D., Makaha Valley Historical Project Summary, Report No. 5. Pacific Anthropological Records No. 31, Department of Anthropology, Bernice P. Bishop Museum, Honolulu, HI.

Handy, E. S. Craighill and Elizabeth Green Handy (1972) *Native Planters in Old Hawaii: Their Life, Lore and Environment*. Bishop Museum Bulletin 233, Bishop Museum Press; Honolulu.

Henry, Jack D., Alan T. Walker and Paul H. Rosendahl (1992) *Archaeological Inventory Survey of Galbraith Trust Lands, Lands of Kamananui and Wahiawa, Waialua and Wahiawa Districts, Island of O´ahu*, Prepared for Helber, Hastert and Fee, Planners, Honolulu, Hawaii.

Huettmann, F. and A.W. Diamond (2001) *Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic*, Ecological Modeling, Volume 141, pgs 261-298.

Kvamme, Kenneth L. (2006) *There and Back Again: Revisiting Archaeological Locational M*odeling in *GIS and Archaeological Site Locational Modeling* edited by Mehrer and Wescott, Pages 3-38.

Lock, Gary and Trevor Harris (2006) *Enhancing Predictive Archaeological Modeling: Integrating Location, Landscape, and Culture*, Chapter 2 in GIS and Archaeological Site Location Modeling, Edited by M. Mehrer and K. Wescott, Taylor and Francis Group, Florida.

Li,An, M. Linderman, J.Qi, A.Shortridge, and J. Liu (2005)*Exploring Complexity in a Human-Environment System: An Agent-Based Spatial Model for Multidisciplinary and Multiscale Integration*, *Annals of the Association of American Geographers*, 1467-8306, Volume 95, Issue 1, 2005, Pages 54 – 79

Mehrer, Mark W., and Konnie Wescott, Editors (2006) *GIS and Archaeological Site Locational Modeling*, Taylor and Francis Group, Florida.

Menze, B.H., J. Ur, and A. Sherratt (2006) *Detection of ancient settlement mounds-Archaeological Survey Based on the SRTM Terrain Model,* Photogrammetric Engineering and Remote Sensing, March 2006.

Mills, E. Nicole (2010) *Analysis of Prehistoric Archaeological Site Distribution Within Fort Campbell Military Reservation, Kentucky-Tennessee, A Thesis Presented to the Faculty of the Department of Geosciences*, Murray State University, Murray Kentucky.

Moore David S., G.P. McCabe and B.A. Craig (2012) *Introduction to the Practice of Statistics*, Seventh Edition, W.H. Freeman and Company, New York.

Parker, Sandra (1985) P*redictive Modeling of Site Settlement Systems Using Multivariate Logistics*, in Archaeological Analysis: Bridging Data Structure, Quantitative Technique, and Theory, edited by Christopher Carr. Kansas City, MO, Westport Publishers, Inc.

Parmegiani, N., M. Poscolien (2003) *DEM Data Processing For a Landscape Archaeology Analysis (Lake Sevan-Armenia), The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol XXXIV, Part 5/W12. http://www.commission5.isprs.org/wg4/workshop_ancona/proceedings/63.pdf

Pierson, Lillian (2017) *Data Science for Dummies*, John Wiley and Sons, Inc., New Jersey.

Roberts, J.J., Benjamin D. Best, Daniel C. Dunn, Eric A. Treml, and Patrick N. Halpin (2010) *Marine Geospatial Ecological Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++*, Environmental Modeling and Sofware, Volume 25, pgs 1197-1207.

Robins, Jennifer J., and Robert Spear (2002) *Cultural Resources Inventory Survey and Limited Testing of the Schofield Barracks Training Areas for the Preparation of a Cultural Resource Management Plan for U.S. Army Training Ranges and Areas, O'ahu Island, Hawaii* (TMK 7-6-01 and 7-7-01). Prepared for U.S. Army Corps of Engineers, Pacific Ocean Division, Fort Shafter, Hawaii by Scientific Consulting Services, Honolulu, HI.

Sanders, Brett. F. (2007) *Evaluation of on-line DEMs for flood inundation modeling*, Advances in Water Resources. Vol. 30, 1831–1843. doi: 10.1016/j.advwatres.2007. 02.005

Sinton, John M., Deborah E. Eason, Mary Tardona, Douglas Pyle, Iris van der Zander, Hervé Guillou, David A. Clague, and John J. Mahoney (2014) Ka'ena *Volcano – A precursor volcano of the island of O'ahu, Hawai'i. Geological* Society of America Bulletin, published online May 02, 2014; doi: 10.1130/B30936.1

U.S. Army Garrison – Hawaii (2018) *An Integrated Cultural Resources Management Plan for the U.S. Army Garrison- Hawaii, Oahu Island* (23April2018), U.S. Army Garrison - Hawai'i.

Williams, Scott S. (2004) Final – *Evaluation of remote Sensing Techniques for Identifying Potentially Significant Cultural Resources at the Makua Military Reservation, Mākua Valley, Island of O'ahu, Hawai'i* (Contract DACA83-95-D-0006, Delivery Order 0017) (February 2004). Prepared for U.S. Army Engineer District, Honolulu, Corps of Engineers, Fort Shafter, Hawaii, prepared by AMEC Earth and Environmental, Inc., Honolulu, HI.

Zeidler, James A. (1995) *Archaeological Inventory Survey Standards and Cost-estimation Guidelines for the Department of Defense*, USACERL Special Report 96/40, Vol 1 December 1995.

(2009) *Final Agency Draft 2009- Integrated Cultural Resources Management Plan 2010-2014 and Environmental Assessment, US Army Garrison - Hawaii (O'ahu and Hawai'i Islands)*, July 2009, The Center for Environmental Management of Military Lands, Colorado State University, Fort Collins, Colorado.

Zeidler, James A., Tania Metcalf, Stephen A. Sherman, Doug Gomez, Marie Held, Catherine Moore, Kai White, Lorna Meidinger, Amanda Murphy, and Alexandra Wallace (2009) Final Agency Draft 2009- *Integrated Cultural Resources Management Plan 2010-2014 and Environmental Assessment*, US Army Garrison - Hawaii (O'ahu and Hawai'i Islands), July 2009, The Center for Environmental Management of Military Lands, Colorado State University, Fort Collins, Colorado.