Prediction to the Max(Net): Comparing Logistic Regression Models to Maximum Entropy Models for Predicting Archaeological Sites within Pitt County North Carolina

By

Elizabeth Wruck

December, 2025

Director of Thesis: Dr. Benjamin Saidel

Major Department: Anthropology

ABSTRACT

Archaeological predictive models are typically used to estimate the likelihood of encountering a prehistoric site within a given area. Traditionally, these models are created using logistic regression, however this requires the use of presence and absence data. This is an issue in archaeology because we work with presence only data. Maximum Entropy models offer an alternative by using presence only data which reduces the uncertainty introduced by randomly generated absence points. This thesis compares logistic regression and maximum entropy (MaxNet) models to determine if there is a difference between their precision and/or accuracy in estimating prehistoric archaeological site locations within Pitt County, North Carolina. Using the same environmental variables, these predictive models were generated for three time periods: a general prehistoric model, an Archaic sites model, and a Woodland sites model. These models were evaluated using P(S|M), precision ratios, and AUC scores. Across nearly all metrics, the Maximum Entropy models preformed as well as or slightly better than the logistic regression models in terms of accuracy. In terms of precision, the Maximum Entropy models preformed significantly better than logistic regression across all metrics. These findings suggest that Maximum Entropy models are on par with logistic regression models in terms of accuracy, and

preform significantly better with precision. These difference in precision can have massive impact on the overall cost of archaeological surveys.

Prediction to the Max(Net): Comparing Logistic Regression Models to Maximum Entropy Models for Predicting Archaeological Sites within Pitt County North Carolina

A Thesis

Presented to the Faculty of the Department of Anthropology

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Arts in Anthropology

By

Elizabeth Wruck

December, 2025

Director of Thesis: Benjamin Saidel, PhD
Thesis Committee Members:
Christopher Wolfe, PhD
Charles Ewen, PhD

Acknowledgements

I have too many people to thank for getting me across this finish line. Thank you, Mom, for always being in my corner and supporting my interests even when you didn’t understand them. Thank you, Stephanie, for always supporting me even when I had lost faith in myself. Thank you to Dr. Charles Ewen for letting me write about this topic even though you had your doubts about it. Dr. Saidel, thank you for stepping in as chair for my committee and making my ramblings legible. And Dr. Wolfe, thank you for helping me make sense of the more difficult math involved in this research. I also need to thank Dr. McMillan for getting me started on this path in archaeology, I would not be where I am today without her guidance in college. I could not have done any of this without the support of my friends, family, and mentors.

Contents

List of Tables

List of Figures

Chapter 1 Introduction

Despite being a discipline that is focused on the study of the past, archaeologists should always be looking for new technological developments because of their potential to improve our investigation of past societies. We can look at other disciplines and take note of what technology and techniques they are developing that might be applicable to our discipline. An example of such borrowing is Geographic Information Systems (GIS), which was initially developed in the 1960's for geographers and is now used by archaeologists nearly as often as their trowels. Looking at similar disciplines and noting relevant methodologies can allow archaeologists to keep up with cutting edge research developing in other fields. In line with this technological advancement, this thesis examines whether maximum entropy predictive modeling could provide advancements in precision and accuracy in locating prehistoric archaeological sites when compared to a logistic regression model.

Predictive modeling has been used within the field of archaeology since the 1980s and is a tool to assess the likelihood of an archaeological site occurring on a landscape. This assessment is based on a set of environmental variables and the known locations of other archaeological sites (Kvamme, 2006, 8). Logistic regression is the statistical technique commonly used when building these models because it is suitable for answering the binary question of a site occurrence; that is, determining the presence or absence of a site, which is the ability to say that we are certain that a site existed at one location and did not exist at another location (Wachtel et al., 2018, 28-29, Yaworsky et al., 2020, 22). Consequently, the need for presence and absence data can be a major weakness when using this model. For instance, the data that archaeologists collect is primarily presence data. This means that archaeologists are able to say that a site existed at one location via the presence of artifacts or archaeological features, but without ground truthing it is impossible to assume that a site never existed at another location. Because archaeological sites are considered rare occurrences on a landscape, researchers are able to get around this issue by randomly selecting points and declaring them as non-sites (Kvamme, 1992). However, this methodology leaves room for error in predicating site location.

This issue has begun to be addressed by archaeologists using techniques borrowed from ecologists to mitigate similar problems in species niche location. A common model utilized in this space is known as the maximum entropy models (Wachtel et al., 2018, 28-29, Yaworsky et al., 2020:22). Maximum entropy models utilize presence only data, meaning that it only requires known site locations and does not utilize randomly generated non-site locations in its prediction. Both Wachtel and Yaworsky have compared logistic regression models against maximum entropy models and found that the maximum entropy models were more effective and efficient when working with archaeological data when the goal is site prediction (Wachtel et al., 2018, Yaworsky et al., 2020). Wachtel et al found that maximum entropy models predicted a higher percentage of the control group while maintaining a higher AUC score than the logistic regression models. For this research I used the MaxNet package in R that was developed by Steven Phillips in 2021 (Phillips, 2021).

Research Question

This thesis is structured around the research question, Are Maximum Entropy models more accurate and/or precise in identifying areas with a higher probability of prehistoric archaeological sites occurring than logistic regression models? Precision in this context is defined by the amount of high probability area generated by the models, while accuracy is the measure of how well the models were able to predict the control data. Using the same data set and environmental variables, I generated a logistic regression and a MaxNet predictive model of prehistoric archaeological sites within Pitt County North Carolina. Additionally, these models were further broken down into three classifications, All Prehistoric Sites, Archaic Sites, and Woodland Sites. A Paleoindian model was not developed because there was not enough site data from that time period within the study area. These models, following the methodology of Wachetel et al. 2018, were then evaluated using their P(S|M), precision ratios, and Area Under the Curve (AUC) scores along with comparing the area of high probability against the area of low probability. P(S|M) measures the probability of correctly measuring site locations within the predicted high-probability areas, precision ratios assess how the model performs relative to random chance by comparing the predicted accuracy to the overall site density, and the AUC score evaluates a models ability to distinguish between site and non-site locations, though this metric is biased towards models that use presence-absence data. My hypothesis is that the Maximum Entropy models will have higher P(S|M), precision ratios, and AUC scores while having smaller areas of high probability when compared to the logistic regression models.

The methodological contribution from this research comes from training and evaluating the logistic regression and maximum entropy models within the same regional dataset. Maximum Entropy models have been tested in other regions in the world; however, it has not been tested within the southeastern United States. Prehistoric sites within this region exhibit variability in material culture and lifeways, ranging from temporary camps to more substantial settlements. Predictive models need to be able to adapt to these changes in lifestyle from the Archaic to Woodland period while still maintaining precision and accuracy.

The outcomes of this research hold direct implication for the cultural resource management (CRM) industry. In large scale surveys, which can oftentimes last multiple years and span multiple counties, utilizing a predictive model for these large-scale surveys allows for these CRM firms to use their time and money more efficiently while still thoroughly surveying areas for cultural material. Small improvements in the precision of these models, especially on a large scale, can lead to a drastic change in the amount of time and labor needed to survey the area. While the improvement of precision is important, it cannot come at the cost of accuracy as this can lead to predictive models missing areas that should be considered as having a high probability of containing prehistoric archaeological sites. That is why the maximum entropy and logistic regression models are being evaluated on both their accuracy and precision. By assessing the differences in accuracy and precision between these two models, we can better assess how effective they would be in phase I archaeological surveys.

Chapter 2 Predictive Models in Archaeology: An Overview

Predictive Models

Archaeological investigations can be broken down into three different phases: I, II, and III, that have different goals and seek to answer different questions (Neumann et al. 2010: 93-108, 135-137, 175-176). A phase I survey aims to locate unknown archaeological sites within a given area. These surveys ask questions such as: If cultural material is found, is there enough to constitute a site? And if there are archaeological sites, with which time period are they associated? Phase I surveys are typically carried out by Cultural Resource Management (CRM) firms that are contracted by land owners to investigate an area to determine if any cultural material would be adversely affected by potential development in an area (Neumann et al. 2010: 93-108).

A phase II assessment is conducted when archaeological sites have been located in an area (Neumann et al. 2010: 135-137). The goal of a phase II assessment is to determine the extent and integrity of the site. At this stage it is also decided whether this site is eligible to be added to the National Register of Historic Places. In phase II the researcher is also seeking to determine how the site fits into the prevailing settlement pattern and it will provide new insights into regional history (Neumann et al. 2010: 135-137).

A phase III data recovery project takes place when a site is eligible for the National Register and preservation is not possible (Neumann et al. 2010: 175-176). The goal of the project is done to recover as much data as possible from the site before it is destroyed. As such, it is sometimes referred to as Salvage Archaeology. An academic Phase III has a specific research goal in mind and is often used to train students, as well as collect data that will be published (Neumann et al. 2010: 175-176). It should be noted that, since the 1980s, nearly all CRM Phase

III investigations are required to address research goals. A Phase III project can be carried by a CRM or academic archaeologist.

Predictive modelling is an underutilized tool that can aid in Phase I work as traditional methods are labor intensive. Typically, this stage uses a combination of pedestrian surveys and subsurface shovel testing, at 15-to-30-meter intervals, in transects (Neumann et al. 2010: 93-108). The process works under the assumption that artifacts within sites are spread out in a sheet, and that the transects will capture most of the site boundaries within the area. In these methods there is room for error as sites can be missed during a survey. Yet, this methodology is often chosen as it is the most practical in terms of time, money and labor (Neumann et al. 2010: 93-108).

Predictive models provide an alternative method for locating archaeological sites by using targeted features, or response variables, derived from independent or predictor variables in a dataset (Lantz 2019, 21). In this project, the target feature is the location of known sites, which are compared against environmental data associated with their occurrence. In archaeology, predictive models of site locations are constructed by identifying environmental factors thought to influence human land-use decisions. These factors are then mapped across a larger area to generate a probability surface, highlighting areas where sites are more or less likely to be found. A probability surface map thus represents zones of high and low likelihood for the presence of archaeological sites (Kvamme 1988, 1990: 261; Schleier 2010, 4). Archaeological predictive models are based on the premise that:

1. Prehistoric people decided where to settle based on environmental and geological factors (Warren and Asche, 2000) (Kvamme, 2006) (White, 1959)

2. These factors still exist on the landscape today (Warren and Asche, 2000) (Kvamme, 2006)

These assumptions weaken the further back in time archaeologists look due to changes in geography and climate, but while they are not perfect, they still serve as a good proxy for the Woodland and Archaic periods within North Carolina. These premises allow for archaeologists to take known site locations and analyze the environmental factors they share (e.g. distance to a water source) and then direct their surveys towards areas that share those environmental factors. Computer-based tools such as ArcGIS and R allow archaeologists to evaluate multiple environmental factors across large areas to develop more complex and more accurate predictive models. ArcGIS Pro is a Geographic Information System (GIS) that was developed by ESRI to manipulate and analyze spatial data (Conolly and Lake 2012: 11). While there are other GIS programs, ArcGIS Pro is the most popular and is what was used for this research. Archaeological data, when in context, has spatial data associated with it that can give insight into cultural patterns or practices. When used for predictive models, spatial data allows for researchers to look for settlement patterns associated with environmental factors. For example, many sites are found close to bodies of water, this is a settlement pattern associated with spatial data. Understanding the lifeways of the groups within the study will be important for interpreting which environmental and geological factors would have been significant to them.

Early Predictive Models

The first wave of predictive modeling in archaeology emerged from processual approaches during the mid-1980s (Kvamme, 2006: 8). One of the landmark works that inspired this first wave was Gordon R. Willey's Prehistoric Settlement Patterns in the Viru Valley, which demonstrated that spatial relationships among sites across a region could reveal broader cultural dynamics at individual locations (Willey 1953S). These first wave models relied on known site locations to estimate the probability of additional sites, drawing on discriminant functions and environmental variables. One example of this is the Glenwood Project, which used canonical discriminant functions (Kvamme, 1980). Canonical discriminant analysis identifies combinations of environmental variables that best separate site from non-site areas. This allows the researcher to evaluate the strength of separation between areas with sites and without sites, to clarify prehistoric land-use patterns. Kvamme developed this model for the Bureau of Land Management (BLM) to be used in place of a predictive map. At the time, predictive maps represented large areas of broad environmental ecotypes like grasslands or forests. Site probability was based on survey sample data associated with each ecotype. Because the Glenwood Project model was developed before geographic information systems (GIS) or personal computers were readily available, the archaeologists would record six environmental variables by hand and enter them into a programmed TI-51 calculator, which produced a p-value estimating the probability of a site's presence based on those variables (Kvamme, 2006: 8).

The second wave of predictive modeling emerged in the mid-1980s with the advent of GIS and a stronger focus on CRM applications. It arose with the availability of Geographic Information Systems (GIS) and was popularized in CRM archaeology (Wheatly and Gillings, 2002: 148). In 2006 Madry et al. developed a GIS-based predictive model of the Piedmont region for the North Carolina Department of Transportation (Madry et al., 2006). The goal of this project was to use the model to plan multi-lane highways throughout North Carolina by ranking proposed highway corridors and alternatives with either a high, medium, or low probability of containing prehistoric archaeological sites. The model used logistic regression to predict archaeological site locations. The predictive model developed by Johnathan Schleier was based on the Piedmont model and is another example of this second wave model. This is relevant because Schleier's model is the only coastal plain model developed for the North Carolina.

While methodologies from second wave models remain popular today, a new wave is on the horizon: ecology-based machine learning methods spearheaded by Wachtel et al.(2018) and Yaworsky et al. (2020). Ecology-based machine learning methods are moving away from regression statistical approaches, which use logistic regression to fit the model using the maximum likelihood estimate, and are instead, using machine learning approaches traditionally designed for species distribution models. These models address the methodological issue of using presence and absence data when creating a predictive model.

Logistic Regression

Regression models test how one or more explanatory variables relate to an outcome variable. For archaeological site location, the explanatory variables are the environmental factors, and a response variable - known prehistoric site locations (Shahbaba, 2011: 253). Archaeologists most often use logistic regression models to predict site locations because it is well suited for situations where the outcome variable is categorical, meaning it falls into a discrete group (Gotelli and Ellison, 2018: 25-26, 273-274). Logistic regression is particularly suited to cases where the response variable is categorical rather than continuous. The explanatory variables are continuous. This means that these variables are quantitative and can take on a value within a smooth interval, like variation in elevation. Categorical variables are qualitative and are used to organize data into meaningful groups. For example, looking at where sites do occur and where they do not occur (Gotelli and Ellison, 2018: 25-26, 273-274).

Logistic regression models estimate the probability of an event—for example, whether a location contains an archaeological site (Kvamme, 1990; Warren, 1990). The model calculates the probability of site presence by applying a logistic function to a weighted sum of the explanatory variables (the environmental variables). The probability ranges from 0 (no site) to 1 (site present) (James, Witten, Hastie, Tibshirani; 2013).

In this context, likelihood refers to how well the model fits the observed data, which is known site locations. The logistic regression model determines the best-fitting relationship between environmental variables and known site location by maximizing the likelihood function. This measures how probable the observed data is with a given set of model parameters. The model can then be applied to surveyed areas to generate a probability map of where archaeological sites have a higher probability of occurring (Kvamme, 1990; Warren, 1990).

In order to generate these models, presence-absence data are needed. This is an issue for archaeologists because most of the data they collect is positive, meaning we know where sites are (positive data) but cannot easily prove where they are not (negative data). The existence of sites can be determined based on surveys, but the absence of data does not mean the absence of a site (Kvamme, 1992). In order to generate non-site data, archaeologists will place random points outside of known site locations and declare them as non-sites (Schleier, 2010). Archaeological sites are seen as rare within a landscape and usually occur in less than 1% of a region. This means that choosing random non-sites should not significantly affect the model (Kvamme, 1992).

This method of picking non-sites leaves room for error when predicting site location (Wachtel et al., 2018; Yaworsky et al., 2020). Unless the area has been previously surveyed and that data has been used to select the non-site locations, there is a chance of an undiscovered site being used to train the non-site model. Additionally, even if historic survey data is taken into account, the survey could have still missed sites, or the sites could have been previously destroyed. The environmental variables in the randomly selected non-site locations can overlap with the environmental variables of the known site locations, which can cause issues in weighing environmental variables. All of this can distort the model and violate the basic assumption that logistic regression uses: sites are present in an area and sites are absent in another area (Wachtel et al., 2018; Yaworsky et al., 2020).

Predictive Models for Eastern North Carolina

The North Carolina Department of Transportation collaborated with archaeologists to develop a predictive model employing logistic regression in seven Piedmont counties to identify areas with a high probability of containing prehistoric archaeological sites and thereby improve highway planning (Madry et al. 2006). Six environmental factors were found to be relevant to site prediction: aspect (north/south), cost distance to water, cost distance to major confluences, cost distance to minor confluences, slope, and topographic variation. Using these factors, seven predictive models were created. The models were designed to test the effect of scope and scale, as well as the varying availability of site data across counties. Two models covered the full seven-county area, two were developed at the county level, one was built at the quad map level, and three used a new technique in which a model for an individual quad incorporated data from surrounding quads to improve accuracy. These models identified areas of high, moderate, and low potential for cultural resources within a 30-meter resolution (Madry et al. 2006).

In his 2010 thesis, GIS Based Archaeological Site Location Modeling in Pitt County, North Carolina, Jonathan Schleier tested the applicability of the Piedmont model in the Coastal Plain. Using 377 previously recorded prehistoric sites from Pitt County, he created four logistic regression models: a Piedmont test model, a Coastal Plain model, a Coastal Archaic model, and a

Coastal Woodland model. Each model was trained on six environmental variables, known archaeological site locations, and randomly selected non-site locations within the county. The models were then evaluated for overall accuracy and for the influence of each environmental variable on the odds of site presence.

Schleier found that most of the Piedmont site-prediction variables were generally applicable to the Coastal Plain. All variables except aspect (north/south) proved significant. The variables were assessed using standard error, Wald statistics, and odds ratios. Ultimately, those that had the strongest impact on predicting site locations in the Coastal Plain were:

I. Topographic Variation	II. Slope	III. Hydric
IV. Cost distance from intermediate stream confluences	V. Cost distance from intermediate/major confluences	

Figure 2-1Environmental variables found to be relevant in predicting archaeological sites within the North Carolina Costal Plain (Schleier, 2010)

He determined this by testing all the environmental variables from the Piedmont model in the Coastal Plain; the details of this process are explained in the methods section. Schleier also found that cost distance to waterways was significant in the Piedmont but not in the Coastal Plain. Instead, Euclidean distance—measured as a straight line between two points—was favored in the Coastal Plain, as opposed to cost distance, which calculates the path of least resistance across a landscape (Schleier 2010). Schleier determined that the most effective variables for the Coastal Plain are soil hydrology (hydric), topographic variation, slope, aspect, distance to intermediate/major waterways, and distance to minor/intermediate confluences. Soil hydrology refers to drainage capacity based on soil type and is grounded in the idea that prehistoric sites were primarily located in areas with well-drained soils. Topographic variation measures changes in elevation using a neighborhood analysis on a Digital Elevation Model (DEM), a type of GIS raster layer. This process “smooths” the image by averaging each cell with those surrounding it, reducing anomalies and emphasizing broader environmental trends (Esri 2024). Prehistoric sites are assumed to occur in areas with mild topographic variation. Slope refers to the rate of elevation change in a DEM and reflects the assumption that sites are typically found on flat or gently sloping terrain. The waterway and confluence variables capture differences in stream size and proximity, which are believed to influence prehistoric site selection (Schleier 2010).

The models developed by Madry et al. (2006) and Schleier (2010) belong to the second wave of archaeological predictive modeling. Both employed logistic regression to build their models. While traditional regression-based approaches, such as logistic regression, rely on maximum likelihood estimation, newer methods adapted from species distribution modeling in ecology address key limitations—particularly the dependence on presence-absence data—through the application of machine learning models.

Chapter 3 The Case Study: An Archaeological and Environmental Overview of Eastern North Carolina

The focus of this research is Pitt County, North Carolina, and this chapter highlights the environment and archaeological record of the state's eastern region. North Carolina's geography is diverse, extending from the Atlantic coast to the Appalachian Mountains. Moving east to west, the Outer Banks form a chain of barrier islands covering roughly 300 square miles along the shoreline. These islands are separated from the mainland by inlets and sounds, which open into the tidewater region's marshy lowlands that extend 30–50 miles inland. Beyond this lies the inner coastal plain—a flat region with some of the most productive farmland in the state, sustained by numerous rivers and tributaries. The inner plain stretches 100–150 miles inland, with less than 500 feet of elevation gain until it meets the fall line. At this point, soft rock gives way to erosion-resistant stone, producing rapids and waterfalls that mark the inland limits of river navigation for large ships (National Geographic 2024).

North Carolina's climate is moderated by the Gulf Stream to the east and the Appalachian Mountains to the west. Average modern temperatures are about 40°F in January and 75.6°F in August (National Centers for Environmental Information 2024).

Pitt County, the focus of this research, is located within the inner coastal plain. It lies in the Middle Atlantic Coastal Plain Ecoregion (Griffith et al. 2002), a landscape characterized by low elevations, flat plains, and numerous swamps, marshes, and estuaries. Soils are typically poorly drained and range from coarse to fine in texture. Two sub-ecoregions occur within Pitt County: the Mid-Atlantic Flatwoods and the Mid-Atlantic Floodplains and Low Terraces. The Flatwoods feature broad upland surfaces, low elevation, limited local relief, and poorly drained soils—contrasting with the more elevated Rolling Coastal Plain sub-ecoregion farther north. The

Floodplains and Low Terraces are associated with large, slow-moving rivers, deep-water swamps, and oxbow lakes (Griffith et al. 2002).

These environmental characteristics have direct implications for the archaeological record. The Tar River and its tributaries provided fertile floodplains that attracted settlement, while poorly drained soils and swampy terrain influenced where communities could live and farm. At the same time, high water tables and alluvial activity affect how sites are preserved, sometimes burying deposits deeply or obscuring them altogether. As a result, the geography of Pitt County not only shaped past human activity but also conditions the ways archaeologists locate, study, and interpret sites today.

Archaeological Overview

The region's ecology and geology present unique challenges. Temporally distinct stratigraphy is difficult to identify because artifacts, such as pottery sherds, migrate vertically in loose sandy soils. High winds further disturb deposits, and the hydrologic cycle causes leaching of organic materials, erasing features like posthole and pits (Ward and Davis, 1999). This is particularly limiting for the Archaic and Woodland periods, where data are already scarce.

Paleo-Indian Period

I will briefly summarize the Paleo-Indian period, even though there are not enough artifacts to include it formally in this study. This overview provides context for the subsequent cultural periods. The Paleo-Indian period, dating to before 8,000 B.C.E., marks the earliest human presence in the Western Hemisphere (Ward and Davis 1999).

Occurring during the Pleistocene epoch, Paleo-Indians were hunter-gatherers who relied on a variety of wild game, including megafauna (Ward and Davis 1999: 32–36). In the western United States, they are understood to have traveled in small groups, primarily hunting mammoths, mastodons, giant bison, and musk oxen, while supplementing their diet with berries, nuts, and small game. In the Southeast, however, environmental conditions supported fewer large herds of megafauna, and Paleo-Indians likely adopted a more generalized subsistence strategy—a precursor to the Archaic pattern.

Temporary shelters were probably made from branches or poles covered with animal pelts. Archaeological evidence from excavated campsites indicates that Paleo-Indians used fire and crafted simple, multi-purpose stone tools. Fluted stone points tipped spears that were thrown with an atlatl. Other implements included knives, choppers, and scrapers for butchering game and processing hides, as well as flint drills and gravers for working bone and wood—materials that rarely survive archaeologically (Wetmore 1975).

In North Carolina, Paleo-Indian sites are identified primarily through Clovis and Folsom spear points, often found as isolated surface discoveries. Dr. David Phelps estimated that during this period the eastern edge of the Coastal Plain lay 230–300 miles east of the Piedmont, suggesting that many Paleo-Indian sites in the Coastal Plain are now submerged underwater (Ward and Davis 1999).

Archaic Period

The Archaic period (8000–1000 B.C.E.) is defined by mobile, non-agricultural cultures that relied on hunting, fishing, and gathering (Ward and Davis 1999). As the Pleistocene epoch ended, the climate warmed to more moderate temperatures, and megafauna became extinct (Ward and Davis 1999). Archaic groups adapted to local conditions by focusing on hunting mammals such as white-tailed deer, black bears, and wild turkeys; fishing and harvesting freshwater and saltwater shellfish; and gathering acorns, hickory nuts, walnuts, seeds, plants, and berries (Ward and Davis 1999; Wetmore 1975). Small bands of extended families or groups of families occupied sites for longer periods, shifting camps seasonally to follow food resources (Ward and Davis 1999; Wetmore 1975).

Many Archaic sites are characterized by large shell middens, reflecting repeated visits over generations, though these are absent in the Piedmont region. Instead, sites such as Doerschuk, Lowder's Ferry, and Gaston contain thick deposits of organically stained soil that reflect intensive occupation (Ward and Davis 1999: 66; Coe 1964: 119). Campsites were widely scattered across the Coastal Plain and Outer Banks, often located near waterways. Streamside camps provided reliable water sources, diverse valley and floodplain resources, and access to natural transportation corridors. These sites can generally be classified as either base camps or temporary procurement sites. Rising sea levels have since submerged many Archaic settlements along the barrier islands, complicating the study of settlement in these areas (Ward and Davis 1999).

Because Archaic groups did not establish permanent villages and had not yet developed pottery, much of the surviving material culture consists of projectile points and lithic debitage. Diagnostic projectile points mark each phase of the Archaic period in the Coastal Plain. Early Archaic sites are identified by Kirk Corner-Notched, Palmer Corner-Notched, Small Hardaway, and Hardaway Side-Notched points (Daniel 2021: 127–154). Middle Archaic sites are associated with Kirk Stemmed, Stanly Stemmed, and Guilford Lanceolate I and II points. Late Archaic sites are marked by Savannah River Stemmed and Thelma Stemmed points (Daniel 2021: 127–154).

As food procurement became more efficient, people devoted more time to craftsmanship. By the Late Archaic period, individuals began to specialize as flintknappers, woodcarvers, or basket-makers, and surplus items were traded for food or other goods. These exchanges fostered long-distance trade networks that circulated raw materials and ornaments made of copper and shell (Wetmore 1975).

Woodland Period

The Woodland period (1000 B.C.E.–1650 C.E.) is divided into three sub-periods: Early (1000–300 B.C.E.), Middle (300 B.C.E.–800 C.E.), and Late (800–1650 C.E.) (Ward and Davis 1999). It is defined by the emergence of agriculture, the development of pottery, the construction of burial mounds, and increasing social complexity (Ward and Davis 1999; Wetmore 1975). Each sub-period is marked by significant cultural and technological change. While pottery did not appear simultaneously across all regions, its presence is used to distinguish Woodland occupations from those of the Archaic period.

Before pottery, food was cooked by stone boiling: fire-heated stones were dropped into skins or baskets filled with water and meat, a slow and inefficient process (Wetmore 1975). The adoption of pottery made cooking more efficient and allowed for improved food storage and transport. Agriculture, in turn, supported larger populations and encouraged more permanent settlement (Wetmore 1975).

The Early Woodland period saw the introduction of pottery, often cord- or fabric-stamped for decoration (Ward and Davis 1999; Herbert 2009). Pottery likely originated in the Savannah River Valley and spread into the North Carolina Coastal Plain through overland and riverine exchange networks (Herbert 2009). In some areas, small villages formed in locations favorable for cultivating specific crops. Although there is no direct evidence of gardening in North Carolina during the Early Woodland, sites in surrounding states show that squash, maygrass, sunflower, chenopod, and sumpweed were being cultivated in small gardens (Ward and Davis 1999).

The Middle Woodland period is associated with population growth and wider settlement, reflected in the increased number of sites and greater material diversity (Herbert 2009). Evidence for gardening becomes clearer, and distinct cultural traditions developed across regions in response to rising populations, economic complexity, and expanding trade networks. By the end of the Middle Woodland, the Outer Banks, Coastal Plain, Piedmont, and Mountain regions were culturally distinct. These groups had different languages and material traditions—for example, Algonkian-speaking groups used shell-tempered pottery, while the Iroquoian-speaking Tuscarora favored pebble-tempered ceramics (Ward and Davis 1999: 210–211).

The Late Woodland period saw an expansion of agriculture, with corn and later beans becoming dietary staples. Increased reliance on farming supported larger populations and led to more complex, permanent villages. Stockades were built around settlements as small- and large-scale tribal conflicts emerged, perhaps driven by competition for farmland, raiding for food surpluses, or ideological differences among culturally diverse groups (Ward and Davis 1999).

Cultural remains from the Late Woodland are more abundant, providing clearer insight into the Coastal Plain's inhabitants. Ceramics from the Inner Coastal Plain resemble those of the Middle Woodland, though tempering materials shifted to small pebbles and fine sand (Ward and Davis 1999). The physical, cultural, and linguistic patterns of this period connect directly to the Algonquian, Iroquoian, and Siouan groups present at the time of European contact.

The shift from seasonal to permanent settlements marks a fundamental difference between Archaic and Woodland land-use strategies. Seasonal mobility and permanent village life involve different environmental requirements, which can complicate predictive modeling. However, in this study such differences should have limited impact, as the selected environmental variables—distance to water, soil hydrology, and slope—are broadly relevant across both periods. To test predictive approaches, three models were created: an all-sites model (combining Archaic and Woodland sites), an Archaic-only model, and a Woodland-only model. This allows for a clearer comparison of how logistic regression and maximum entropy perform across different cultural contexts and provides more precise models for each period.

Chapter 4 Predictive Models for Pitt County Model: The Methodology

All of the Pitt County predictive models were created using archaeological site data provided by the Office of State Archaeology. At the time of this research, there were a total of 463 prehistoric sites located within Pitt County. Of these sites, 146 dated to the Woodland period, 165 were from the Archaic period, and 10 the Paleo-Indian period, 139 sites were unable to be identified to a specific prehistoric period and were just labeled as prehistoric. It is important to note that some sites were designated as both Woodland and Archaic, due to multiple periods of use. While the Archaic and Woodland time period have different settlement patterns, hunter-gatherer vs farming, the environmental variables used within the models would be relevant to deciding where to settle across both time periods. In general people settled in areas that were flat, close to water, and did not easily flood. This is consistent with the environmental variables that Schleier found to be effective to predicting both Archaic and Woodland site locations within Pitt County (Schleier, 2010).

The hydric environmental layer (Fig 4-1) contains data from the Soil Survey Geographic Database (Soil Survey Stall, 2024). The slope layer (Fig 4-2) was built using a digital elevation model (DEM) that was obtained from the NC OneMap website. The Hydrology environmental layer was built using data from the National Hydrography Dataset (National Hydrography, 2024).

Environmental Layers

The environmental data were compiled based on the variables that Johnathan Schleier found to be significant in his 2010 thesis. Schleier concluded that there were five variables that were effective for archaeological predictive modeling within the coastal plain: hydric, topographic variation, slope, intermediate/major waterways, and minor/intermediate waterways. Though I attempted to recreate the intermediate/major waterways, and minor/intermediate waterways, I was unable to recreate the environmental layers based on their description within Schleier's thesis. This could be due to changes within ArcGIS since the previous study was done or it could be from my own user error. Instead, I replaced those layers with areas within 100 meters of water and areas within 300 meters of water.

When working with geographic information it is important to have a general understanding of projection systems. A map projection is a way to display areas of the Earth's round surface onto a flat map (Conolly and Lake, 2012; 17-18, 2-23). This involves mathematically transforming units of longitude and latitude into a two-dimensional space. At a large, continental, scale this can cause distortion and spatial errors. For regional mapping projects, such as this one, US State Plane systems are often a better choice because they often maximize spatial accuracy for the specific area covered by that system. All the data should be converted into the same coordinate system to ensure that data points are accurately placed geographically in relation to each other. (Conolly and Lake, 2012; 17-18, 22-23). All environmental layers were converted to the North Carolina State Plane NAD 1983 coordinate system.

Hydric

The hydric environmental layer (Fig 4-1) contains data from the Soil Survey Geographic Database (Soil Survey Stall, 2024). The vector for Pitt County was converted into a raster within ArcGIS in order to allow for proper data extraction. The .CSV file that contained the soil information then had to be attached to the raster. A vector in ArcGIS uses lines and points to represent locations while a raster uses cells to represent locations, like pixels in an image (ESRI, 2018). Raster were used for the model due to the large amount of data that was being analyzed. Additionally, because of the large scale of this model, the cells were able to still accurately represent areas within Pitt County. The .CSV file that contained the soil information was attached to the raster.

The most pertinent information for this research was in the drclassdcd column of the attribute table which described how well drained the soils were. This column divided the soils into five categorical values: poorly drained, somewhat poorly drained, moderately well drained, well drained, and excessively drained. I then created a column called *hydriclev* where I categorized the data into three drainage types and assigned each type an integer value (Table 4-1). I included a no data category to account for any areas within the map that did not contain any data.

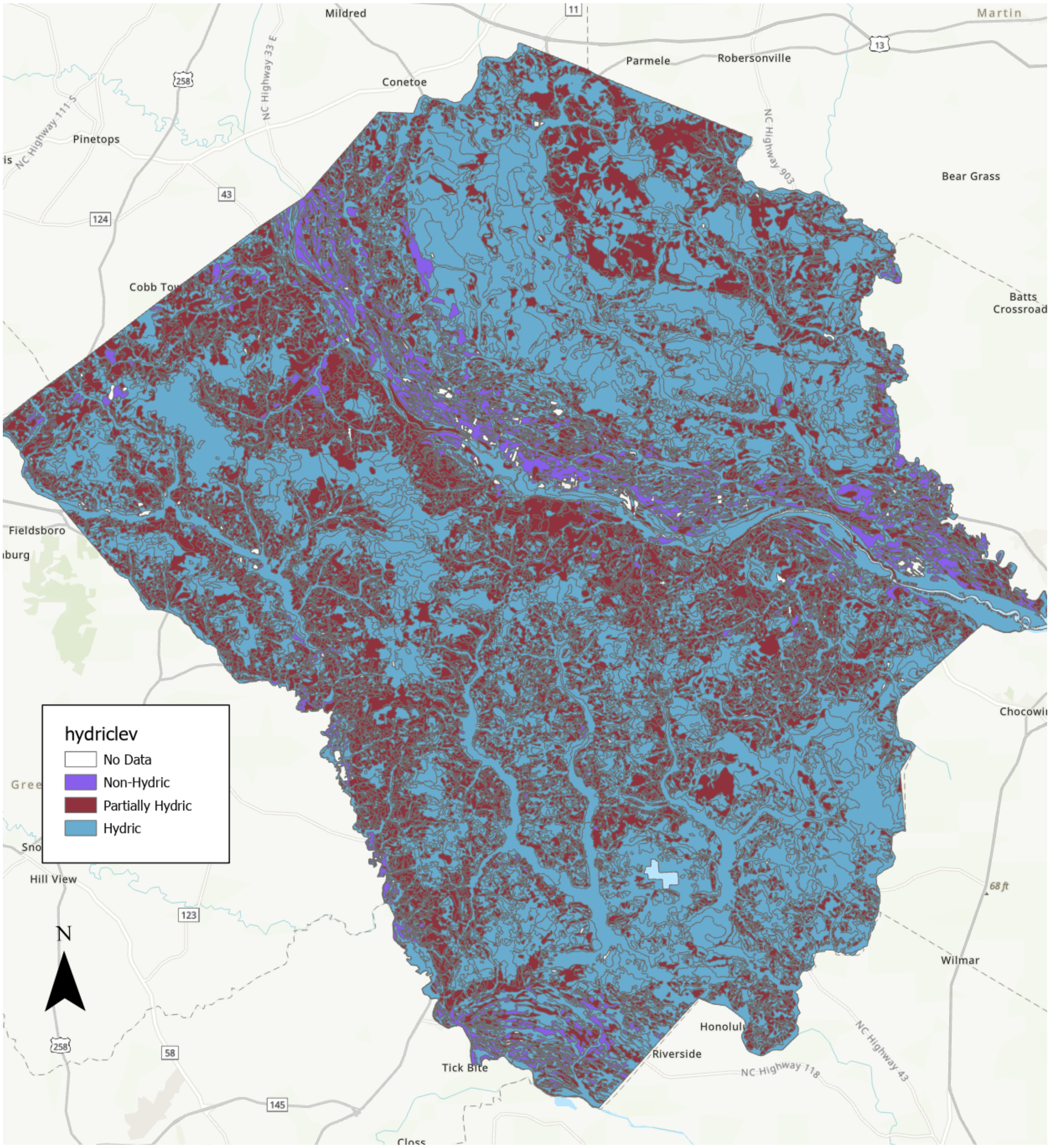


Figure 4-1Map representing the different soil hydric level within Pitt County, NC.

Classification Number	Soil Type	Soil Class
1	Non-Hydric	Excessively Drained
2	Partially Hydric	Well Drained & Partially Well drained
3	Hydric	Poorly Drained & Somewhat Poorly Drained
0	No Data	Contains no Class data

Table 4-1 Soil Classification

Slope

The slope layer (Fig 4-2) was built using a digital elevation model (DEM) that was obtained from the NC OneMap website. I applied a neighborhood analysis to the DEM, which is a spatial statistics tool that allows analyses of a cell by using the cells around it. This ‘smooths’ the image by averaging the cell with those around it, which reduces the effects of anomalies within the data and focuses on broader trends within the environment (Esri, 2024).

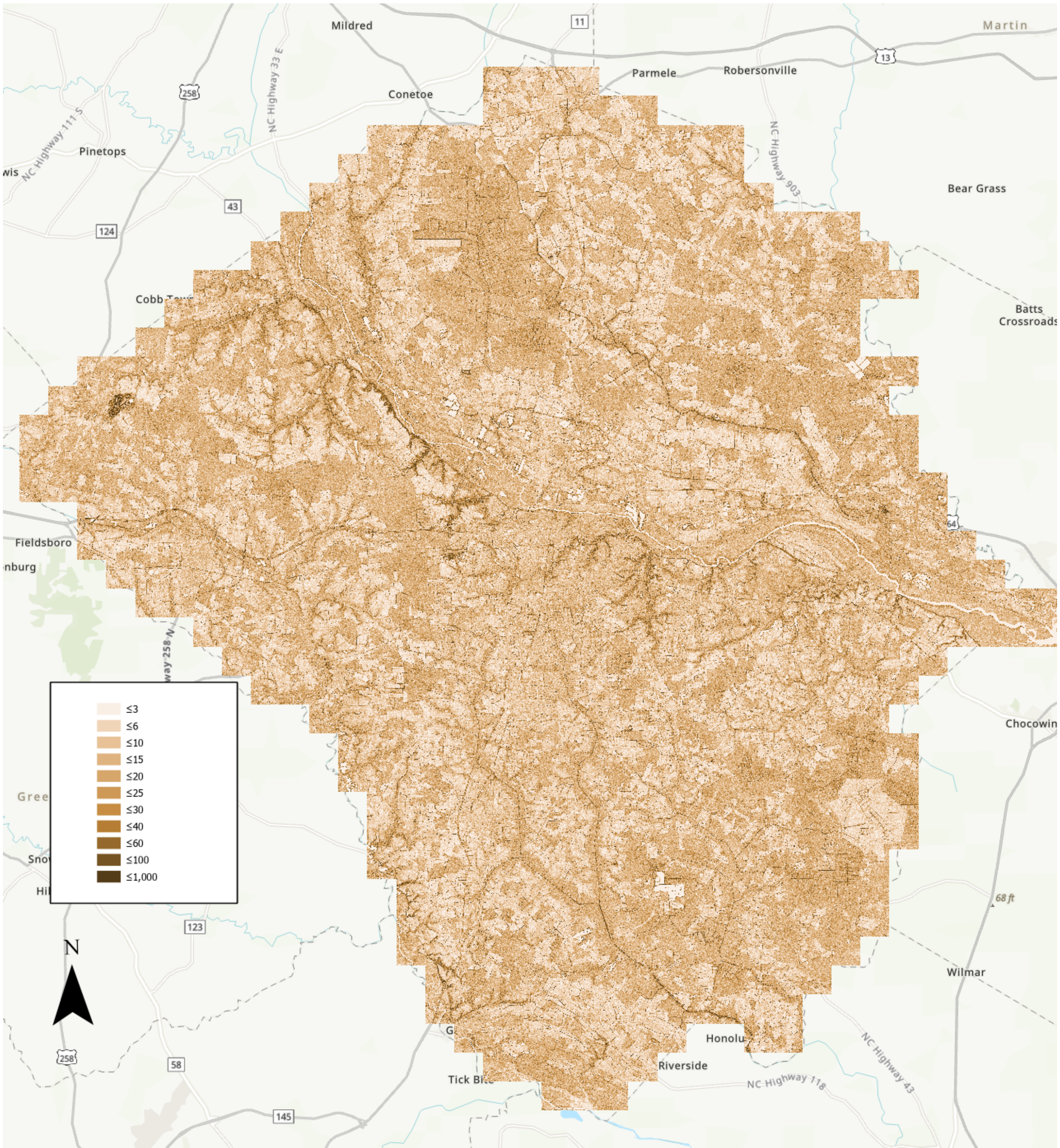


Figure 4-2 Map representing the slope across Pitt County, NC.

Hydrology

Since I was unable to replicate Schleier's hydrology layer, I utilized a tutorial created by Dr. Edward Gonzalez-Tennant to create two different environmental layers: area within 100 meters of water and area within 300 meters of water (Gonzalez-Tennant, 2020). In order to generate these layers, I downloaded data from the National Hydrography Dataset (National Hydrography, 2024). I then used the buffer tool to generate a buffer area 100-meters and 300-meters around any water drainage networks within Pitt County.

Machine Learning

Machine learning is a computational approach that enables models to identify patterns and make predictions based on data. Unlike traditional statistical methods, machine learning relies on iterative processes to refine models by minimizing errors or loss functions. These models can adjust parameters, often starting with simple or random values and gradually optimizing them through repeated testing (Lantz, 2019: 21, Yaworsky et al. 2020).

There are two main subgroups within machine learning: supervised learning, and unsupervised learning (Lantz, 2019: 21). Unsupervised learning uses unlabeled output data, and the model independently identifies patterns or relationships within the data. Predictive models use supervised learning methods to train the model. This means that the model works to optimize a function by finding the combination of independent variables that result in an observed output (Lantz, 2019: 21). With archaeological data, that means the model analyzes all the combinations of environmental factors that have been input and finds a pattern of those factors that occur at archaeological sites. So, when developing areas of high and low probability, the model will look for areas that match the combinations seen at the known site locations.

Machine Learning can be divided into a five-step process: data collection, data exploration and preparation, model training, model evaluation, and model improvement (Lantz, 2019: 18). Data gathering consists of collecting the data points that will be used to train the model. For this thesis, it involved collecting the archaeological sites location data from the NC Office of State Archaeology, and the environmental layers for Pitt County. Data preparation involves cleaning the data so that it is suitable for the model. This involved converting the environmental data into the specific variables that were tested for in the model, removing non-relevant data, and converting the layers into .CSV files that could be used within *R*. *R* is a coding language and program that is used for statistical computing and graphics (The R Foundation).

Model training refers to the process of fitting the predictive model to the dataset by adjusting its parameters to identify patterns between independent variables, the environmental data, and the dependent variables, archaeological site location. For logistic regression, this is done by estimating the regression coefficients that best predict site presence. In MaxEnt and MaxNet, the training is done by repeatedly adjusting how the environmental variables are weighted to maximize entropy while fitting the site data provided. This is the step in the process where the map of high and low probability areas will be produced. Model evaluation is the analysis of how well the machine learning program conducted its task. Model improvement entails editing methodology when better performance from the model is required (Lantz, 2019: 18).

Maximum Entropy

MaxEnt is a machine-learning model that was developed by Steven Philips, Miro Dudik, and Rob Schapire in 2004 to generate species distribution models (SDMs) and find species niches within ecology (Phillips, Dudik, and Schapire, 2004: 1). This program is run through the open-source software *R,* which is a coding language used for statistical analysis (R-Project). MaxEnt uses the principle of maximum entropy to make uniform and constrained predictions. Entropy is a concept within machine learning that quantifies randomness or uncertainty within a group of values referred to as a class. A low entropy value indicates a homogeneous data set, while a high entropy value represents a diverse data set (Lantze, 2019: 132). MaxEnt models start with the assumption that all locations in the area of interest are equally likely to contain archaeological sites. The model then adjusts the site probability based on the environmental variables provided (constraints) (Merow, C., Smith, M.J., Silander, J.A. 2013). This differs from logistic regression which looks at the environmental variables at known sites locations and randomly selects non-site locations and creates the site probability based on matching environmental variables at known sites and randomly selected non-sites. In the context of archaeology, the key difference between logistic regression and maximum entropy models is that the maximum entropy models will only use the location of known archaeological sites to make decisions on patterns of environmental variables that line up with known site location.

Machine learning models approximate the relationship between species' sightings and the environment in which those sightings occurred (Franklin, 2009). Because species' relationship with the environment are complex it is best to represent that relationship with a nonlinear function (Austin, 2002). MaxEnt uses the presence-only data of species sightings along with environmental predictors across a predetermined area and compares it to locations where presence is unmeasured, allowing them to predict the species niche (Merow, C., Smith, M.J., Silander, J.A. 2013). Logistic regression models require the use of true absence points, which does not align with the presence-only data that archaeology produces. MaxEnt uses pseudo-absence points, which are selected at random within areas that presence is unlikely to occur in locations where presence has not been observed (Merow, C., Smith, M.J., Silander, J.A. 2013). By using pseudo-absence points, MaxEnt addresses the issue of randomly selected non-site data by selecting locations where presence is unlikely, rather than assuming true absence. This approach allows for a more accurate estimation of site probability by better reflecting the nature of archaeological data collection (Wachtel et al., 2018; Yaworsky et al., 2020).

Yaworsky et al. (2020) explain the relevance of MaxEnt to archaeological data. MaxEnt begins with the assumption that all locations are equally as likely to have a site which is a distribution of maximum entropy, or the most uniform distribution that still fits the data. The model then takes each environmental variable and forms a probability density function (Fig 4-3) where the height or density of the function is proportional to how often that variable occurs within the landscape. Predictor variables then increase the density of values that are at known archaeological sites, which constrains the distribution. This allows variables to be equally weighted, so environmental factors that occur in larger portions of the study area will have the same probability as factors that occur in smaller portions of the study area. The model uses the assumption that a site is equally likely to occur on a landscape, (i.e., a distribution of maximum entropy), to create a uniform distribution and the constraints work using environmental variables to create a non-uniform distribution. A less biased model is created that retains the most uniform distribution within the constraints. This type of model is ideal for archaeological sites as it accounts for a larger range in variation and uses pseudo-absence points (Yaworksy et al., 2020). In simple terms, MaxEnt assumes that archaeological sites are equally likely across the study area, and then adjusts that assumption based on the environmental variables (like soil type and distance to water) to match where known sites are located. This creates a balanced model that avoids bias or using random data, like non-site locations.

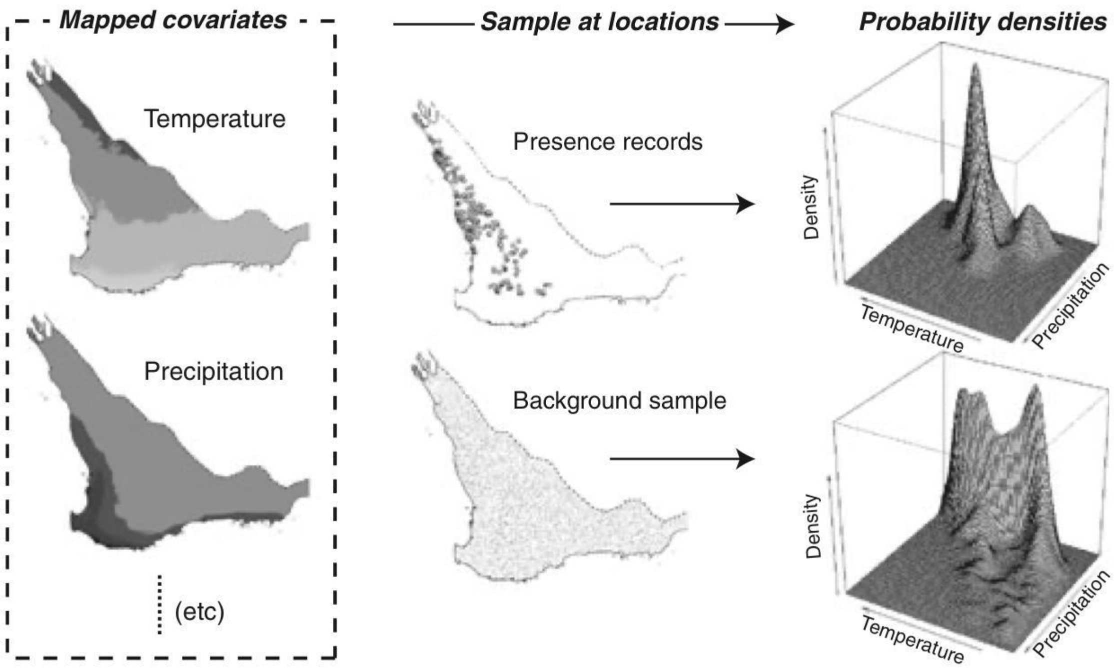


Figure 4-3 A diagram representing probability densities for Species Density Models using MaxEnt. The maps to the left show distribution of environmental factors (covariates). The middle shows the location of presence and background samples. The map to the right shows the distribution of values within covariate space. The bottom right sample shows the distribution of the covariates within the background sample, and the top right map shows the distribution of covariates with the presence data. Courtesy of Elith et al 2010 "A Statistical Explanation of MaxEnt for ecologists" a Journal of Conservation Biogeography

In two studies comparing predictive models (Wachtel et al., 2018; Yaworsky et al., 2020) MaxEnt was found to be the most accurate and efficient for archaeological data. Wachtel et al. (2018) compared logistic regression models and maximal entropy models in independent locations in both north Israel and northeast China. They found that both models produced useful data, however, the MaxEnt model outperformed logistic regression in terms of predictive accuracy and model robustness. Specifically, in both study areas, MaxEnt predicted a higher percentage of archaeological sites while having a smaller high probability area. MaxEnt also worked better with smaller data sets, when using only five known sites for the training data, MaxEnt predicted 68.6% of the control group with a high probability area of 7.1% of the total project area. With the same training data, logistic regression was unable to generate a usable model for 5 out of the 20 runs, and on average it predicted 45.4% of the control group with the high-probable area covering 45.9% of the total project area. The AUC scores were higher for MaxEnt, 0.94, when compared to the logistic regression scores, 0.78-0.83 (Wachtel et al., 2018).

Yaworksy et al. (2020) compared four statistical approaches within predictive modeling: binomial logistic regression, a generalized additive model, Maximum Entropy (MaxEnt), and Random Forests. This study was conducted in the Grand Staircase-Escalante National Monument in Utah with Formative Period (2100-650 BP) sites. They found that while Random Forests had the highest AUC score in the model evaluation, this evaluation was biased because it assumed that pseudo-absence points are true-absence points. They found that MaxEnt, the model that used pseudo-absence points, is most suitable for archaeological data (Yaworsky et al., 2020).

While logistic regression is still a useful tool for hypothesis testing, these initial reports indicate that MaxEnt provides better results for predictive models for archaeological sites. The uniform distribution and constraints set up within MaxEnt work well with archaeological data as it is often non-uniform and has large variations in environmental factors. Additionally, the use of pseudo-absence points allows for there to be no error generated from false absence points that are used within logistic regression models.

MaxNet was published in 2021 by one of the creators of MaxEnt, Steven Phillips (Philips 2021). MaxNet uses the same model structure as MaxEnt, but employs the 'glmnet' package to fit the model. Model fitting is used in data science to adjust a model's parameters to most accurately match the model's predictions against the data that were input (Lantz 2019, 12-17). Glmnet is a software package created in *R* that is used to build different types of statistical models. Glmnet uses penalized maximum likelihood to best fit the model to the data. Maximum likelihood estimation is a method that estimates the parameters of a model, but this method can lead to overfitting the model. Penalized maximum likelihood incorporates a regularization framework to balance goodness of fit with the model complexity. This helps prevent overfitting by variance and bias, which leads to a model that generalizes better to new data. This improves convergence and stability (Friedman et al., 2023). This works in a very similar way to MaxEnt which uses uniform distribution against constraints to generate the least biased model with the most uniform distribution within those constraints. MaxNet does not need Java, another coding language, to generate the model. This allows for fewer resources being used by the computer and less room for error due to the fact that only one coding program is being used, *R*, instead of two, *R* and *Java*. Thus, the program can be run on smaller computers like laptops. The MaxNet package is what was used in this thesis, it is important to reiterate that the MaxNet package runs a maximum entropy model that is the same as the MaxEnt package, its differences are in how it is fit and that it is only run through *R*.

Code

The code for the predictive model was written in *R*. Prior to this thesis I had no background in R and only minor experience in coding. I used *Chat GPT-4,* along with textbooks, as a tool to look over my code when I hit errors. I would check my code by copying and pasting the code and the error message into the *Chat GPT-4* and typing commands like 'Where is the error and explain step by step how to fix it.' I was able to get simple explanations on how to fix my code. There was some trial and error in using this method, where the solutions I was provided were not correct, however I was able to quickly realize the error when I ran the code. Checking and editing the code with *Chat GPT-4* was done throughout the entire process of building the code. It is important to note that archaeological site location data was not uploaded to *Chat GPT-4* due to its sensitive nature. More information on how *Chat GPT-4* was used will be detailed in the discussion chapter.

The code was written using *R Markdown*, which is a file format where large sections of the code are knitted within the document. This allows for inclusion of plain text so that the code can be annotated, this ultimately creates an html document of the code with step-by-step explanations of what the model is doing (Grolemund, 2014).

The first step when coding in *R* is to load the libraries which will attach packages to the program. These packages contain codes and commands that have been created and published by other users. This allows for the code to be more concise and legible. For example, the 'dismo' package contains tools for species distribution modeling. This is where the *maxent()* command is pulled from. Rather than attempting to build a machine learning model from scratch, the dismo package already contains the code to build the MaxEnt model and other tools for preparing the data and evaluating the model (Hijmans, Phillips, Leathwick, and Elith; 2023). Packages in *R* are published with a guide that includes all of the commands within the package and explanations of what they do. The packages that were used within my code were: pRoc (Robin et al., 2011), dismo (Hijmans et al., 2024), raster (Hijmans R, 2025), and data.table (Barrett et al., 2025).

The second step in the code was to set the raster options. This was a single line of code that divided the rasters into smaller chunks. This was required because the spatial analysis of rasters is a memory intensive task. Breaking the rasters into smaller pieces allowed 14 of the CPUs within the computer to analyze different sections of the raster images concurrently. This made the code run faster and be less memory intensive.

The data were then loaded and prepared in the code. The environmental layers that were prepared in ArcGIS were converted into *.tif* files and loaded into R. The logistic regression model requires non-site data. This section of the code generated random points within the bounds of the environmental layer and then created a table listing the point ID, longitude, and latitude, mimicking the format of the site data. The next step was to extract the raster values. Site data were assigned a value of 1 while non-site data were assigned a value of 0.

The next section of the code combined the site data and the background points into the same data table. The table was then shuffled and split into two groups, a model training dataset which represented 70% of the total site data and a model testing data set which was made up of the remaining 30% of the total site data. The training data set uses the location of site, and for the case logistic regression, non-site data to recognize patterns of site locations. The testing data were used to evaluate the model, since these sites were excluded from training it, they can be compared against the predictive model to analyze its accuracy.

Once the environmental and site location data are prepared, the next step was divided into two parts: creating the Logistic Regression model and creating the MaxNet model.

The Logistic Regression model is trained using a data table that combines site and non-site data with the corresponding environmental variables. Based on the data provided, the model learns the relationship between the environmental variables and where sites do and don't occur. Once the model is trained, it is able to predict the likelihood of a site occurring at a given location based on the environmental variables present there. Though at this point there is no image or plot. The model was fitted with a binomial family and a logit link.

For the MaxNet model, the training data was prepared for the model, this involved removing columns for the data table that were used to label the data, specifically the Long (longitude), Lat (latitude), and Pres (site presence indicated with 0 and 1 values) labels. This became the predictor variable. The training_data was then reduced to just data points where sites occurred, this became the response variable. The MaxNet model was then run using both the response and predictor variables.

Evaluation

The models were evaluated using Wachtel et al.'s (2018) methodology to test and compare the effectiveness of their logistic regression and MaxEnt models. There were three methods used to evaluate the predictive models: P(S|M), precision ratio, and Area under the Curve (AUC).

P(S) represents the likelihood of an archaeological occurring at any point in a study area, the formula to calculate it is:

$$P(S) = \frac{Total\ Area\ of\ Sites}{Total\ Study\ Area}$$

while M represents the total area of the high probability area (Wachtel et al., 2018, Kvamme 2006: 24-29). P(S|M) gives the true positive rate and represents the probability of locating an archaeological site within a given model. It evaluates the relationship between the model's high probability areas and the known site locations. This tells you how accurately the model was able to predict archaeological sites within high probability areas, the closer to 1 the more accurate the model is at predicting where high probability areas are (Wachtel et al., 2018, Kvamme 2006: 24-29). This was used to measure and evaluate the accuracy of the model while using presence only data.

The precision ratio is an index of model fit (Wachtel et al., 2018, Kvamme 2006: 24-29). The formula to calculate it is:

$$\frac{P(S|M)}{P(S)}$$

The precision ratio shows how well the model is predicting sites in relation to how common they are in the data. A higher ratio means that the model is more effective at identifying high probability areas and that these high probability areas are more likely to identify sites when comparing the P(S) score of the study area. This score represents the probability of an archaeological site occurring when the model specifies a high probability area. If the score is higher than 1, the model is better at identifying sites than random chance, if the score is equal to 1 the model performs equally to random chance, and if the score is less than 1 the model performs worse than random chance (Wachtel et al., 2018, Kvamme 2006: 24-29).

The Area under the Curve (AUC) score was used by both Wachtel et al., 2018 and Yaworsky et al., 2020, to evaluate their models. The AUC uses a curve to plot the true positive rate against the false positive rate. The AUC shows how well a model can predict compared to random guessing by plotting the true positive rate against the false positive rate. This generates a score between 0 and 1 where the closer to 1 the more accurate and specific the model is (Lantz 2019: 332-331). A model is considered accurate if 80%-85% of the observations from the control group are within the high-probability areas of the model (Lantz 2019: 332-331). It is important to note that because the AUC uses presence and absence data to generate its score, it will be biased towards models that use presence and absence data, like logistic regression (Yaworsky et al., 2020).

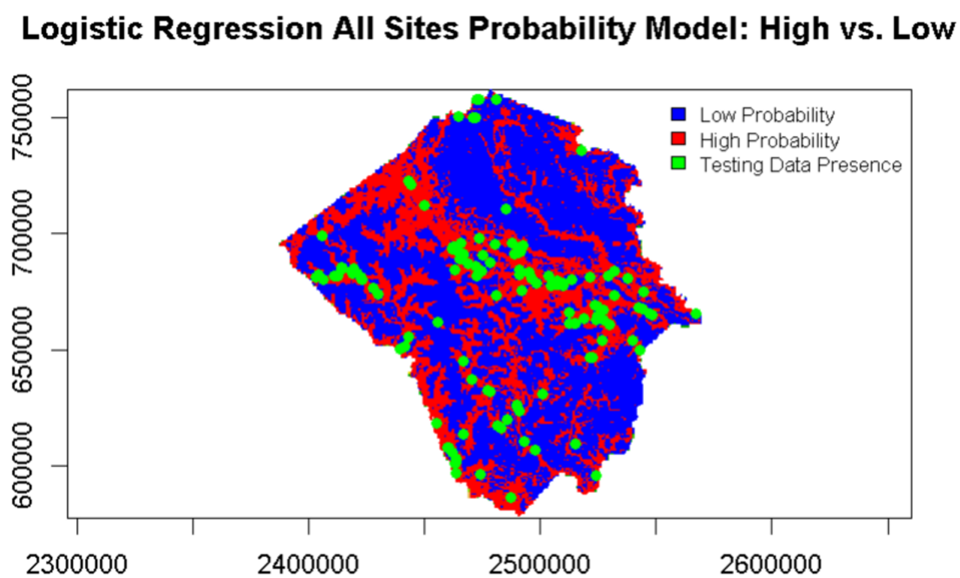
Chapter 5 Results

In summary MaxNet received the highest scores across all three evaluation methods for the all site model and the Archaic model. For the Woodland model, MaxNet received the highest scores for the P(S|M) and precision ratio while the logistic regression received the highest score for the AUC. It is important to restate that the AUC is biased toward the logistic regression model as it evaluated based on presence and absence data (Yaworsky et al., 2020). A more in-depth discussion of the results follows.

The results for my research will be presented in three sections. The first discusses the models that used all the site data. The second discusses the models that were created using only Archaic site data. The third section will focus on models that were created using Woodland site data. All these models were created using the same code and environmental variables: slope, soil drainage, areas within 100 meters of water, and areas within 300 meters of water.

All-sites Model

The all-sites model was created using the 463 documented prehistoric sites in Pitt County. From a visual standpoint, the MaxNet model has a smaller high probability area than the logistic regression model (Fig. 5-1). This is confirmed when comparing the precision ratio scores, with MaxNet having the highest score. The model with the highest P(S|M) and AUC scores was MaxNet (Table 5-1). All of the models score's fall withing their respective 95% confidence intervals which indicates that they have stable performance under resampling. The MaxNet values for P(S|M) and the precision ratio are closer to the upper bounds of their confidence intervals which indicate that the MaxNet model preforms slightly better than the logistic regression model in identifying areas with true site presence. Overall, MaxNet had the highest score across all three metrics for the all-sites model. This indicates that the MaxNet models preform slightly better than logistic regression model in those metrics.



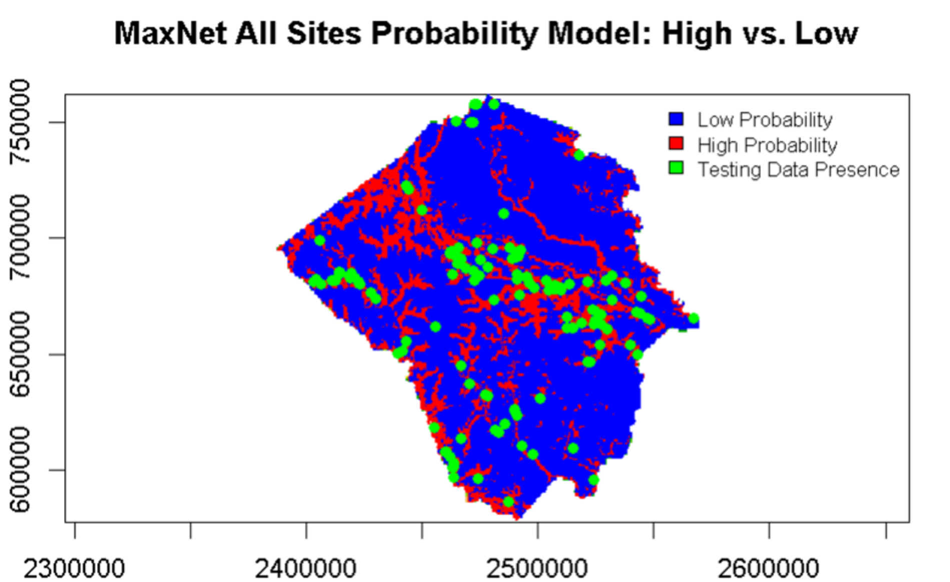


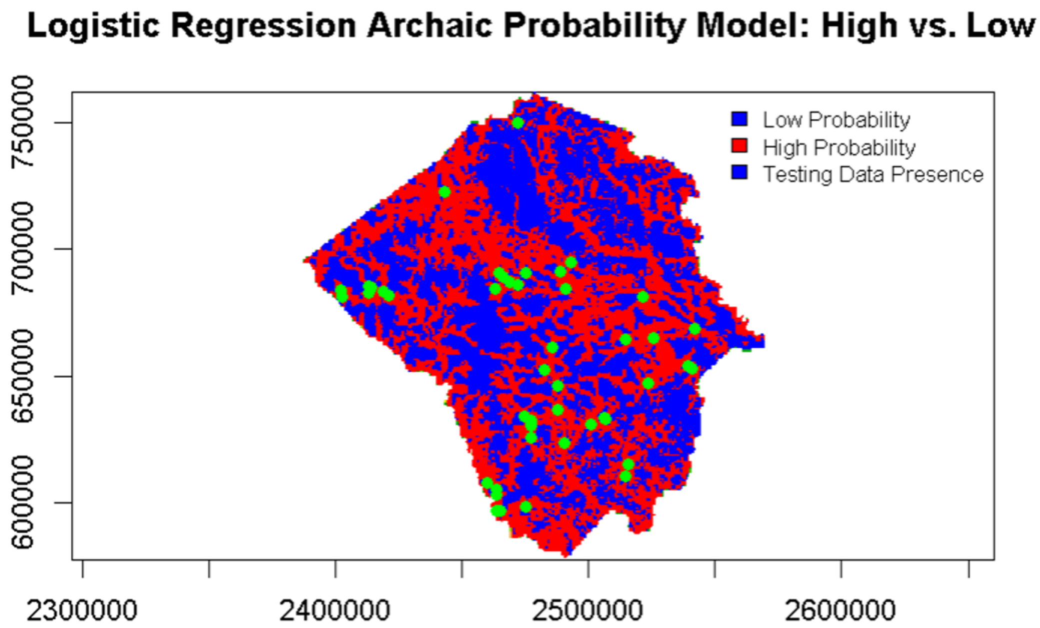
Figure 5-1Logistic Regression and MaxNet probability models of all prehistoric sites in Pitt County, North Carolina. Low probability areas are represented with blue, high probability areas are represented by red, and the green dots represent the site locations use

	P(S|M)	Precision Ratio	AUC
Logistic Regression	0.7533 *(0.682 – 0.826)*	1.2744 *(1.187 – 1.377)*	0.7802 *(0.718 – 0.841)*
MaxNet	**0.7982** *(0.724 – 0.872)*	**1.3504** *(1.241 – 1.497)*	**0.7913** *(0.726 – 0.851)*

Table 5-1 The evaluation Scores for the All Sites models with 95% bootstrapped confidence intervals

Archaic Sites Model

The Archaic sites models were created using the 167 documented Archaic sites in Pitt County. From a visual standpoint, the MaxNet model has a smaller high probability area than the logistic regression model (Fig. 5-2). This is confirmed when comparing the precision ratio scores, with MaxNet having the highest score. The model with the highest P(S|M) and AUC scores was MaxNet (Table 5-2). All of the scores for the evaluation methods fall within their confidence intervals, with the MaxNet scores for the P(S|M) and precision ratio being closer to the upper limit. This indicates that while both models show stable and reliable results, the MaxNet model has marginally stronger predictive performance and precision than the logistic regression model.



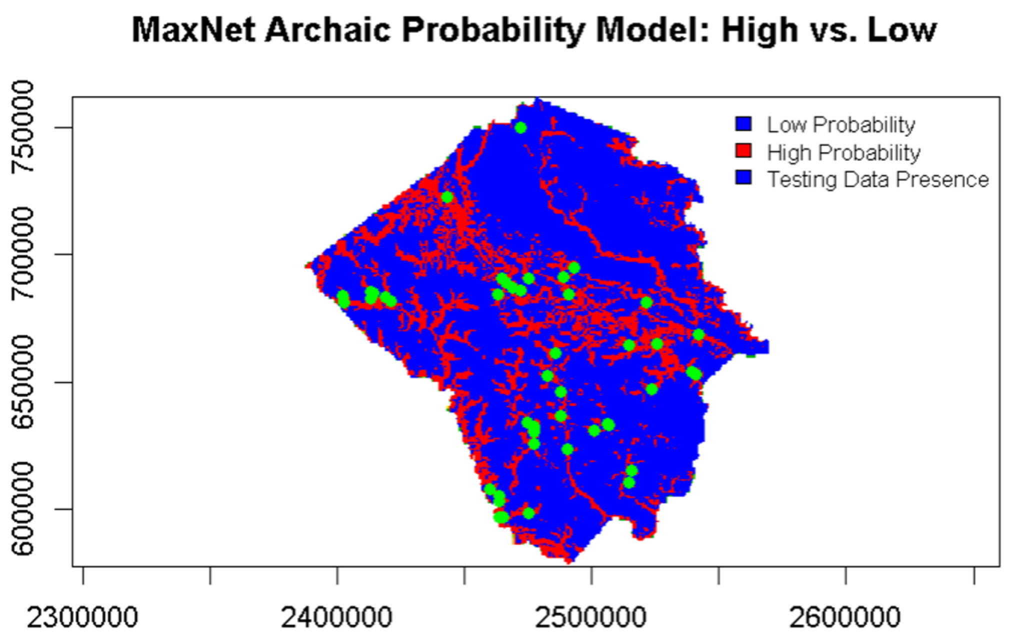


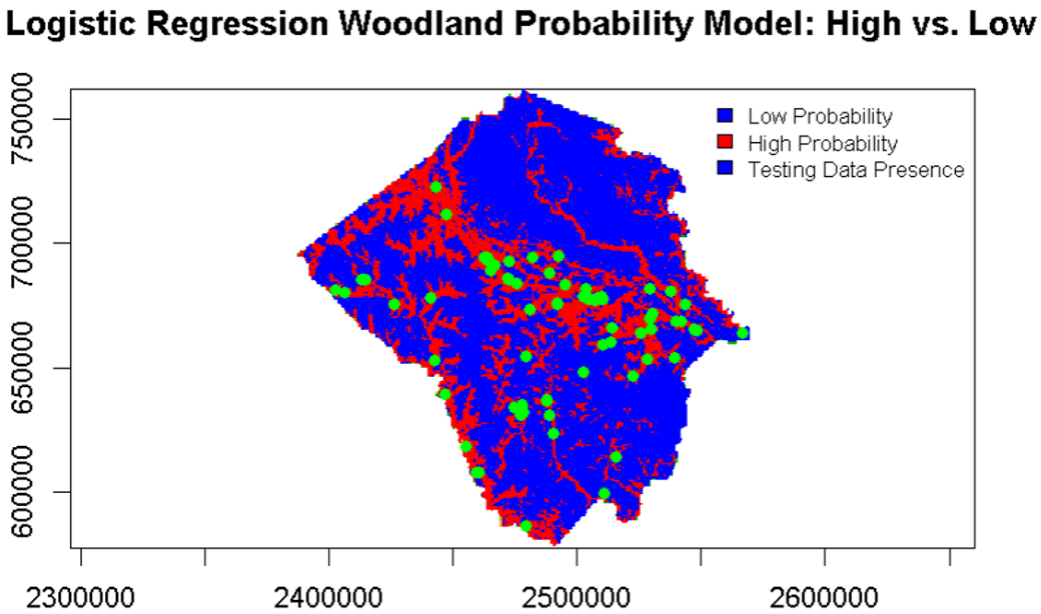
Figure 5-2 Logistic Regression and MaxNet probability models of Archaic sites in Pitt County, North Carolina. Low probability areas are represented with blue and high probability areas are represented by red and the green dots represent the site locations

	P(S|M)	Precision Ratio	AUC
Logistic Regression	0.7368 *(0.621 – 0.845)*	1.3483 *(1.191 – 1.454)*	0.8096 *(0.699 – 0.899)*
MaxNet	**0.8055** *(0.666 – 0.933)*	**1.4740** *(1.242 – 1.780)*	**0.8297** *(0.701 – 0.902)*

Table 5-2 Evaluation scores for the Archaic Sites models with 95% bootstrapped confidence intervals

Woodland Sites Model

The Woodland Sites models were created using the 238 documented Woodland sites in Pitt County. Visually, the MaxNet model only has a slightly smaller high probability areas than the logistic regression model (Fig 5-3). This is confirmed when comparing the precision ratio scores, with MaxNet having the highest scores. When looking at the P(S|M), MaxNet scored the highest. Additionally, the logistic regression model has the highest AUC score and based on the confidence intervals, preforms marginally better than the Maxnet model (Table 5-3). It should be noted that these are the highest AUC scores from the study. Overall, the MaxNet model has marginally stronger predictive precision for the woodland sites while the logistic regression model has meridianally stronger predictive accuracy for the woodland sites.



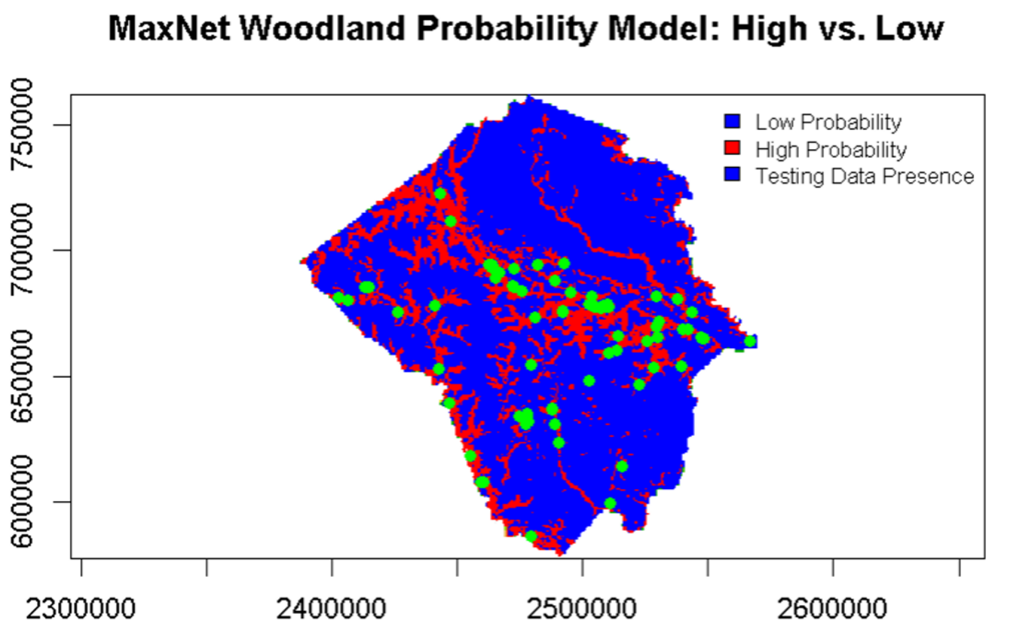


Figure 5-3 Logistic Regression and MaxNet probability models of woodland sites in Pitt County, North Carolina. Low probability areas are represented with blue and high probability areas are represented by red and the green dots represent the site location

	P(S|M)	**Precision Ratio**	**AUC**
Logistic Regression	0.8784 *(0.797 – 0.946)*	1.3869 *(1.248 – 1.569)*	**.8854** *(0.812 – 0.950)*
MaxNet	**0.9245** *(0.840 – 0.983)*	**1.4598** *(1.292 – 1.681)*	0.8696 *(0.785 – 0.939)*

Table 5-3 Evaluation scores for the Woodland Sites models with 95% bootstrapped confidence intervals

Chapter 6 Discussion

When comparing all three models across all evaluation metrics, maximum entropy most consistently produced the highest scores. Logistic regression achieved the highest value in only one instance—the AUC score for the Woodland Sites model. Although the differences between the scores are modest, the results indicate that maximum entropy performs at least as well as, and in some cases slightly better than, logistic regression in predicting archaeological site locations. Both the P(S|M) and precision ratio metrics rely on presence-only data, meaning the models are evaluated on their ability to predict high-probability areas based on known site locations.

When evaluating model performance using the Area Under the Curve (AUC), it is important to note that this metric relies on presence–absence data. As a result, AUC values tend to favor methods such as logistic regression that incorporate presence–absence inputs. Similar results were reported by Yaworsky et al. (2020). Visual inspection of the predictive models with testing site locations plotted (Fig. 5-1–5-3) shows that predicted site locations correspond closely with high-probability areas across all models. The primary difference is that the maximum entropy model produces a smaller high-probability area. The only notable exception is a single site in the southwest corner of the study area: the maximum entropy model did not classify this location as high probability, whereas the logistic regression model did.

The primary advantage of using maximum entropy over logistic regression is that it produces smaller high-probability areas while still maintaining strong predictive performance. To compare the extent of high- and low-probability areas across the all-sites models, I converted the raster cells to square miles in R and calculated the number of cells falling within each probability category. I then plotted these values in a bar chart (Fig. 6-1).

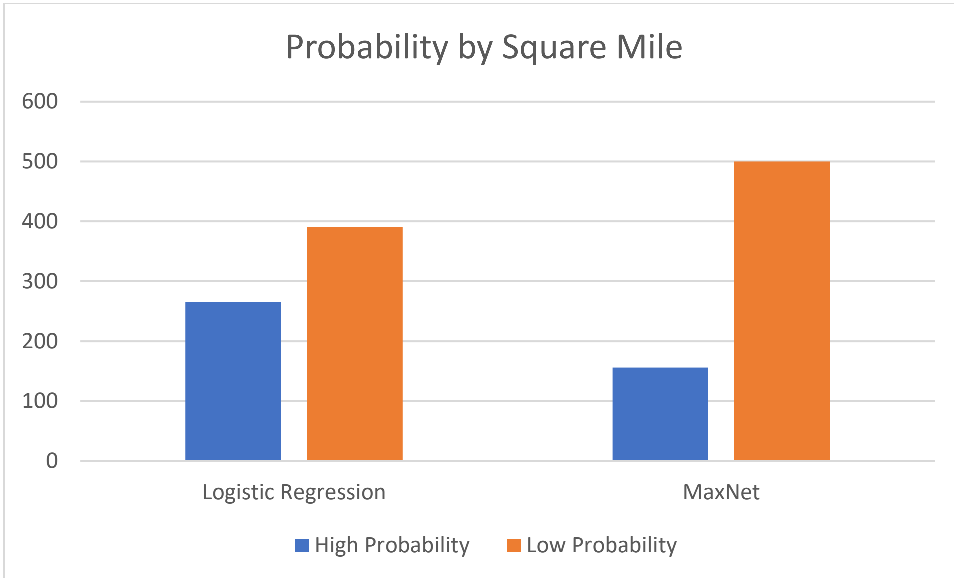


Figure 6-1 Bar Graph showing the square mile of high Probability areas vs low probability areas of the All Site models.

Additionally, I converted the high-probability areas for each model into percentages representing the proportion of the predictive map classified as high probability (Fig. 6-2). This comparison demonstrates that the MaxEnt and MaxNet models yield smaller high-probability areas than the logistic regression model.

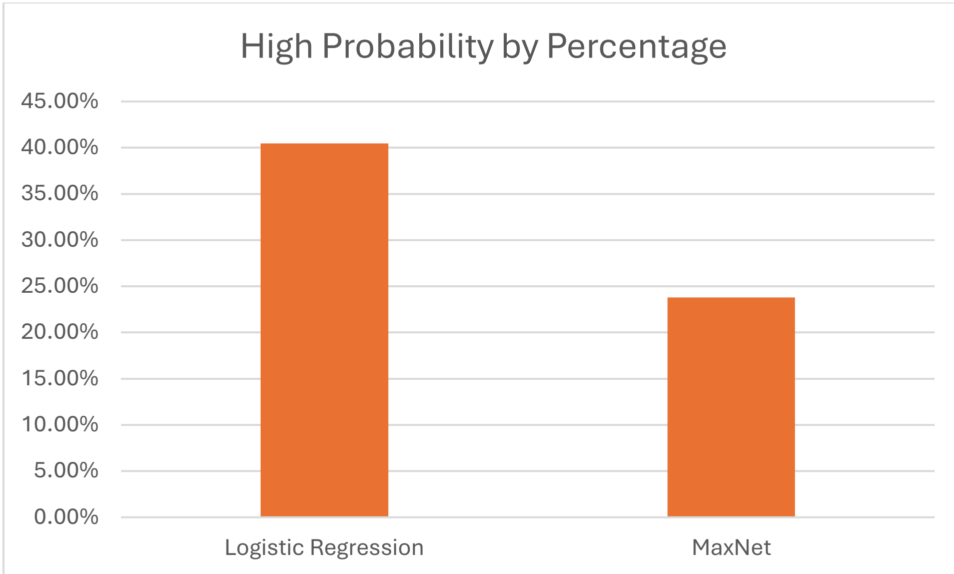


Figure 6-2 Bar Graph showing the High Probity by percentage of the area

Overall, while the AUC scores indicate that all three models have similar predictive accuracy, the primary difference lies in their precision. When viewed with the understanding that AUC is biased toward presence–absence models, it is notable that all scores fall within 0.02 of one another. Lantz (2019) explains AUC score interpretation as being similar to academic grading:

A	Outstanding	0.9 to 1.0
B	Excellent/Good	0.8 to 0.9
C	Acceptable/Fair	O.7 to 0.8
D	Poor	0.6 to 0.7
F	No discrimination	0.5 to 0.6

Figure 6-3 A recreation of Lantz 2019's AUC score interpretation based on academic grades.

Because the thresholds between AUC performance categories span large ranges—for example, AUC values between 0.8 and 0.9 indicate an excellent or good model—very small differences in

AUC values, such as those observed here, do not meaningfully change how the models are evaluated (Lantz 2019:333). Since all AUC scores in this study are within 0.002 of each other, and given the metric's bias toward presence–absence models rather than presence-only models, the results indicate that the models exhibit comparable predictive power. In contrast, their spatial precision differs more noticeably. The maximum entropy model produced high-probability areas that were approximately 15% smaller than those generated by the logistic regression model, indicating higher precision. This pattern is supported by the precision-ratio scores, which show that the MaxNet models had a higher likelihood of encountering archaeological sites within their high-probability areas. Taken together, these results suggest that MaxNet provides a more refined prediction, reducing the extent of high-probability area while still maintaining predictive accuracy.

There are too many variables to accurately estimate the cost of conducting an archaeological survey across all of Pitt County, particularly given factors such as private land ownership, roads, and urban development. Instead, I calculated the percentage of high- and low-probability areas generated by both the logistic regression and MaxEnt models across a hypothetical 1,000-acre area to estimate the cost difference between the two approaches. The total cost was calculated using an assumed cost of $50 per shovel test pit (STP). Additionally, this estimate assumes that high-probability areas are surveyed at 15-meter intervals and low-probability areas at 30-meter intervals.

	Logistic Regression	Maximum Entropy
STPs in High Probability area	107,920	64,752
STPs in Low Probability area	80,940	102,524
Total STPs	188,860	167,276
Total Cost	$9,443,000	$8,363,800

Table 6-1 Table showing how many STPs are estimated to be within the high and low probability areas of 1000 acre area with the same percentage of high and low probability areas as the logistic regression and maximum entropy models.

Based on these calculations, there is a $1,079,200 difference in cost resulting from the number of STPs required. This estimate does not include additional expenses such as per diem, lodging, and other logistical costs that would also increase with additional field time. This substantial difference underscores how model precision can significantly affect the overall cost of archaeological fieldwork.

Having smaller high-probability areas allows archaeologists to use predictive models more efficiently. While logistic regression is a simpler method, the MaxEnt model's ability to generate smaller high-probability areas justifies its use for archaeological site prediction. However, increased precision also introduces a potential trade-off, as narrower high-probability zones may increase the risk of missing archaeological sites during survey. Balancing efficiency with comprehensive coverage is therefore essential, and future applications of these models in CRM contexts should consider strategies to mitigate the risk of false negatives while still benefiting from improved spatial precision.

In CRM, both accuracy (correctly identifying the presence of archaeological sites) and precision (reducing the area of high probability while still detecting site locations) are critical. Overly broad high-probability areas can increase labor costs and extend project timelines, whereas overly narrow areas may result in archaeological sites being missed. This is why it is significant that the MaxEnt and logistic regression models produce comparable accuracy. These models can help reduce the time and resources required for construction projects by allowing archaeologists to more precisely assess the likelihood of encountering prehistoric sites in a given area. Additionally, they enable cultural resource management firms to allocate their workforce more efficiently during surveys while still thoroughly investigating high-probability zones

Chat GPT

ChatGPT-4 is incorporated into this thesis as a tool to assist with generating and editing code in R. ChatGPT is a publicly accessible large language model (LLM) that has been widely adopted within the academic community as a tool for supporting scientific writing (Oguz, Ekersular, Sunnetci et al., 2023). LLMs are trained on large datasets and are designed to summarize and generate text (Backström and Kihlert, 2023). These tools have been shown to be useful in educational settings because they enable students to receive personalized learning support tailored to their needs and preferences (Kasneci et al., 2023). Because this is an emerging and continually developing technology, it is important to recognize its limitations. LLMs can generate inaccurate or fabricated information (Meyer and Urbanowicz, 2023). When prompted to answer a question for which they lack sufficient knowledge, these systems may produce false responses rather than acknowledging uncertainty.

I used ChatGPT primarily to generate and edit R code and verified its responses against external sources. A study conducted in Sweden evaluated whether ChatGPT could improve the code quality of novice programmers and found that it helped identify and remove code-quality violations (Backström and Kihlert, 2023). My only prior coding experience was an introductory Java course taken in my junior year of high school. I incorporated ChatGPT-4 as a tool to help me write the R code for this research. For example, I would prompt the program with requests such as, “Write R code to compare logistic regression predictive archaeological site location models to maximum entropy predictive models, and explain the code step by step,” and receive the following output:

> Certainly! Below, I will walk you through the steps of comparing Logistic Regression models to Maximum Entropy models for predicting archaeological site locations using R.
>
> ## Step 1: Loading Necessary Libraries
>
> R
> Copy code
> ```
> install.packages(c("glm", "dismo"))
> library(glm)
> library(dismo)
> ```
>
> ## Step 2: Load and Prepare the Data
>
> R
> Copy code
> ```
> # Assuming your data is in a CSV format
> data <- read.csv("your_data.csv")
>
> # Divide data into sites (1) and non-sites (0)
> sites <- data[data$site == 1, ]
> non_sites <- data[data$site == 0, ]
>
> # Split the data into training and testing datasets
> set.seed(123) # for reproducibility
> training_rows <- sample(1:nrow(data), 0.7 * nrow(data))
> training_data <- data[training_rows, ]
> testing_data <- data[-training_rows, ]
> ```
>
> ## Step 3: Create a Logistic Regression Model
>
> R
> Copy code
> ```
> logistic_model <- glm(site ~ predictor1 + predictor2 + ... , data=training_data,
> family="binomial")
> summary(logistic_model)
> ```
>
> ## Step 4: Create a Maximum Entropy Model using MaxEnt
>
> R
> Copy code
> ```
> # Prepare occurrence and background points
> occurrences <- sites[sample(1:nrow(sites), 0.7 * nrow(sites)), ]
> background <- non_sites[sample(1:nrow(non_sites), 0.3 * nrow(non_sites)), ]
> ```

```
# We use only predictors in this example
predictors <- stack(predictor1, predictor2, ...)

# Fit the MaxEnt model
maxent_model <- maxent(predictors, p=occurrences, a=background)
```

Step 5: Predict and Evaluate Models

R
Copy code

```
# For Logistic Regression
logistic_predictions <- predict(logistic_model, newdata=testing_data, type="response")
logistic_preds_binary <- ifelse(logistic_predictions > 0.5, 1, 0)

# For MaxEnt
maxent_predictions <- predict(maxent_model, predictors)
maxent_preds_binary <- ifelse(maxent_predictions > 0.5, 1, 0)

# Compare accuracy
logistic_accuracy <- sum(logistic_preds_binary == testing_data$site) /
length(logistic_preds_binary)
maxent_accuracy <- sum(maxent_preds_binary == testing_data$site) /
length(maxent_preds_binary)

print(paste("Logistic Regression Accuracy:", logistic_accuracy))
print(paste("MaxEnt Accuracy:", maxent_accuracy))
```

Explanation:

1. **Step 1:** Libraries for logistic regression (glm) and MaxEnt (dismo) are loaded.
2. **Step 2:** The data is loaded and split into training and testing sets. It's assumed that sites are labeled with 1 and non-sites with 0.
3. **Step 3:** A logistic regression model is created using specified predictors.
4. **Step 4:** A Maximum Entropy model is fitted using the maxent() function, specifying occurrence points and background (or absence) points.
5. **Step 5:** Predictions are made for both models on the testing dataset, and accuracy is compared.

Note: You'll need to replace predictor1, predictor2, ... with your actual predictor variable names and ensure that "your_data.csv" represents the correct path to your data file.

This simple outline was extremely helpful because it provided the building blocks for code that I could refine and expand. Additionally, ChatGPT-4 was useful when I encountered errors; I could copy and paste my code and the error message, and the model would identify issues—sometimes a simple spelling mistake, other times suggesting methods for debugging more complex problems. I used ChatGPT-4 much like a tutor that reviewed my code and offered guidance on how to proceed. While tools like ChatGPT are not required to write the code used in this research, they make the modeling process more accessible for individuals without a strong coding background. It is also important to note that I supplemented this assistance with traditional resources, such as R textbooks and online programming forums, to troubleshoot and better understand the code. This approach ensures transparency in the use of artificial intelligence tools and maintains academic integrity by combining automated assistance with independent verification and traditional learning resources.

The most challenging aspect of coding the models was ensuring that the environmental layers and site data were formatted correctly so the program could read them. Writing the logistic regression code was relatively straightforward and required minimal data manipulation. In contrast, the MaxEnt model was more difficult to implement without prior experience and required additional data preparation to function properly. MaxEnt also relies on a Java package, meaning the code was executed through both R and Java. This dual-language process introduced more opportunities for errors that could cause the program to crash and required the use of a high-powered desktop computer.

While troubleshooting MaxEnt, ChatGPT recommended using MaxNet instead. After further research, I found that MaxNet produces models comparable to MaxEnt while being easier to implement and less computationally intensive. The package was developed by Steven Phillips, one of the original creators of MaxEnt, and one of its main advantages is that it is fully integrated into the R ecosystem and does not require Java (Phillips 2022). MaxNet utilizes the glmnet package to fit models, applying a penalized likelihood approach to approximate maximum entropy rather than calculating it directly. In other words, instead of computing MaxEnt directly, glmnet uses regularization to prevent overfitting while estimating a distribution similar to MaxEnt. This method requires considerably less computing power while still producing models that are comparable to—if not more accurate than—the logistic regression and MaxEnt models, while also offering greater spatial precision (Phillips 2022). Most importantly, MaxNet was straightforward to implement in R and still produced high-quality results.

While there is a learning curve associated with using R to build predictive models, the process becomes more manageable when supported by tools such as ChatGPT. In the future, applications similar to Wallace could be developed for archaeological modeling. Wallace provides a user-friendly interface for the MaxEnt framework, allowing users to input site data and environmental layers without writing code. It is widely used by ecologists to create and visualize species distribution models, which is the purpose for which MaxEnt was originally designed (Kass et al., 2017; Kass et al., 2018). Making similar tools accessible to archaeologists would help integrate predictive modeling into routine practice within the discipline. Such developments would enable cultural resource management practitioners to apply predictive modeling more efficiently during project planning and survey design, ultimately improving archaeological outcomes while reducing time and cost burdens for compliance work.

Chapter 7 Conclusion

The research question examined whether maximum entropy models are more accurate and precise than logistic regression models in identifying high-probability areas for prehistoric archaeological sites. The results indicate that both models produced similar levels of accuracy based on their AUC scores. It is important to note, however, that AUC scores are biased toward presence-absence models. When evaluating P(S|M) and precision ratio scores, maximum entropy models only slightly outperform logistic regression. Both metrics have limitations, but the findings suggest a marginal advantage for maximum entropy.

Overall, the evaluation methods show that maximum entropy models are slightly more precise than logistic regression. Although maximum entropy requires a steeper learning curve, these models address methodological challenges associated with presence-only site location data. Logistic regression, however, performs comparably to maximum entropy while being easier to implement and requiring less training. Additional research comparing these modeling approaches in archaeological contexts is necessary, though existing studies are promising (Wachtel et al., 2018; Yaworsky et al., 2020). The hypothesis that maximum entropy would produce higher P(S|M), precision ratios, and AUC scores—and smaller high-probability areas than logistic regression—is partially supported. With the exception of the woodland model's AUC score, maximum entropy consistently produced slightly higher scores and smaller high-probability areas while maintaining similar accuracy. Despite these encouraging results, several limitations must be acknowledged. AUC scores are biased toward presence-absence data; P(S|M) offers only a point estimate and may obscure variability across iterations; and precision ratios ignore false negatives and may be sensitive to class imbalance. Furthermore, this analysis is based on a single study area within one county. Additional research is needed to verify these initial findings

Recommendation for Future Research

Logistic regression remains an effective tool for predicting prehistoric archaeological site locations. This study, however, suggests that maximum entropy models perform equally well and yield smaller high-probability areas. Maximum entropy performed best on the Archaic dataset, which had the smallest sample size, indicating that it may be particularly useful when data are limited. Future research should therefore examine model performance across a wider range of sample sizes. Additional studies focused on the U.S. East Coast would expand the growing literature on predictive modeling in archaeology, and a larger Coastal Plain model—similar to the existing Piedmont model—should be developed. One ongoing barrier to broader adoption of maximum entropy by archaeologists is the need for coding proficiency. Developing a user-friendly interface, comparable to the Wallace platform used in ecology, would facilitate wider use, especially since R is free and accessible. Future work should also explore how tools such as ChatGPT-4 and similar large language models can support archaeologists in implementing and interpreting these predictive modeling approaches.

Pennsylvania currently uses a statewide predictive model to guide archaeological survey work. This system adjusts shovel test pit intervals based on whether an area falls within high-, moderate-, or low-probability zones (PA State Historic Preservation Office). As a result, CRM firms can allocate personnel more efficiently while still meeting professional survey standards. In contrast, the North Carolina State Historic Preservation Office (NC SHPO) currently relies on ArcMap, a legacy GIS program that will no longer be supported after March 2026 (North Carolina Office of State Archaeology; Esri ArcMap Life Cycle). The upcoming software transition presents an opportunity for NC SHPO to adopt a predictive modeling framework similar to Pennsylvania's. A probability-based approach would allow archaeologists to focus on high-probability areas more intensively while reducing unnecessary testing elsewhere, saving time and resources. As digital technologies evolve, archaeologists must remain current with new tools and methods. Although coding may seem like a barrier, emerging software interfaces and AI-assisted tools are making predictive modeling increasingly accessible across the discipline

As technology continues to evolve, it is essential for archaeologists to adopt new analytical tools and workflows. Coding requirements can appear to be a significant barrier, particularly for practitioners trained before the widespread adoption of R and Python in archaeological research. However, advances in user-friendly interfaces and AI-supported coding assistance are lowering these barriers. Tools such as R-based graphical platforms, no-code GIS applications, and large language models capable of generating executable scripts provide opportunities to integrate predictive modeling into archaeological practice more easily. By embracing these emerging technologies, archaeologists can improve analytical rigor, increase efficiency, and remain aligned with methodological developments in related disciplines such as ecology, geography, and data science.

Continued investment in training and professional development will be critical for successful implementation of these approaches. Universities, state agencies, and CRM firms should prioritize workshops, short courses, and collaborative partnerships that introduce practitioners to predictive modeling, open-source GIS tools, and coding fundamentals. Integrating these skills into graduate curricula and internship programs will help prepare the next generation of archaeologists to work effectively with emerging technologies. Importantly, these training initiatives should emphasize practical and applied exercises tied to real CRM and research needs, ensuring that methodological innovation translates into improved field strategies, better resource management decisions, and stronger archaeological outcomes.

Bibliography

Austin, M.P.

2002 Spatial Prediction of Species Distribution: An Interface between Ecological Theory and Statistical Modeling. Ecological Modelling, 157, 101-118.

Backström, O., & Kihlert, A.

2023 Code Quality and Large Language Models in Computer Science Education : Enhancing student-written code through ChatGPT (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-330999

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B, Krylov I

2025 data.table: Extension of 'data.frame'. R package version 1.17.99, https://r-datatable.com.

Coe, Joffre L.

1964 The Formative Cultures of the Carolina Piedmont. Transactions of the American Philosophical Society, n.s., 54, pt. 5. Philadelphia, Pa.: American Philosophical Society.

Conolly, James; Mark Lake

2012 Geographical Information Systems in Archaeology. Cambridge University Press, Cambridge, UK.

Elith J, Phillips SJ, Hastie T, et al.

2011 A statistical explanation of MaxEnt for ecologists. Divers Distrib 2011; 17: 43–57.

Enkelejda Kasneci et al.

2023 "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103, p. 102274. issn: 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274. url: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

Esri

Hydrogeology Toolset, ArcGIS Pro. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-flow-accumulation-works.htm Accessed September 19, 2024.

Esri

Neighborhood Summary Statistics (Spatial Statistics), ArcGIS Pro. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/neighborhood-summary-statistics.htm. Accessed September 19, 2024.

Esri

2019 Raster is faster but vector is corrector. Esti.com, https://www.esri.com/content/dam/esrisites/en-us/media/pdf/teach-with-gis/raster-faster.pdf. Accessed September 19, 2024.

Esri

ArcMap Life Cycle: ArcGIS Desktop 10.8.2. https://support.esri.com/en-us/products/arcmap/life-cycle. Accessed December 2024.

Feng, Xiao; Cassondra Walker, Fikirte Gebresenbet.

2017 A Brief Tutorial on Running Maxent in R. Github.com, https://github.com/shandongfx/workshop_maxent_R/blob/master/code/Appendix1_case_study.md. Accessed April 13, 2023.

Franklin, J.

2009 Mapping Species Distributions: Spatial Inferences and Prediction. Cambridge University Press, Cambridge, UK.

Fiedman, Jerome; T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, and N. Simon

2023 glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. (R package version 4.1-4) https://cran.r-project.org/web/packages/glmnet/index.html

Gonzalez-Tennant, Edward

2020 Processing Steps for a Simple Archaeological Predictive Model. https://anthroyeti.net/QGIS4Arch/SimplePredictiveModelingSteps.pdf. Accessed September 19, 2024.

Gotelli, Nicholas J.; Arron M. Ellison

2018 A Primer for Ecological Statistics, Second Edition. pp.25-26 and 273-274. Oxford University Press, New York.

Griffith, G.E., Omernik, J.M., Comstock, J.A., Schafale, M.P., McNab, W.H., Lenat, D.R., MacPherson, T.F., Glover, J.B., and Shelburne, V.B.

2002 Ecoregions of North Carolina and South Carolina, (color poster with map, descriptive text, summary tables, and photographs): Reston, Virginia, U.S. Geological Survey (map scale 1:1,500,000).

Grolemund Garrett

2014 Introduction to R Markdown, R Markdown from R Studio. https://rmarkdown.rstudio.com/articles_intro.html

Hastie, Tevor; Juyang Qian, and Kenneth Tay

2023 An Introduction to glmnet. https://cran.rproject.org/web/packages/glmnet/vignettes/glmnet.pdf

Herbert, Joseph M.

2009 *Woodland Potters and Archaeological Ceramics of the North Carolina Coast*. The University of Alabama Press, Tuscaloosa.

Hijmans, RobertJ.

2025 raster: Geographic Data Analysis and Modeling. R package version 3.6-32, https://rspatial.org/raster.

Hijmans, Robert J; David R. Carpenter; J.B. Phillips; Et al.

2023 dismo: Species Distribution Modeling (R package version 1.3-6) https://cran.r-project.org/web/packages/dismo/dismo.pdf

James, Gareth; Daniela Witten; Trevor Hastie; Robert Tibshirani;

2013 An Introduction to Statistical learning: with Applications in R. Springer, New York. (Pg 130-131)

Kass JM, Pinilla-Buitrago GE, Paz A, Johnson BA, Grisales-Betancur V, Meenan SI, Attali D, Broennimann O, Galante PJ, Maitner BS, Owens HL, Varela S, Aiello-Lammens ME, Merow C, Blair ME, Anderson RP.

2023 wallace 2: a shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. Ecography, Volume 2023(3), e06547. https://doi.org/10.1111/ecog.06547

Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C, Anderson RP.

(2018) *Wallace*: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution*, 9, 1151–1156. https://doi.org/10.1111/2041-210X.12945

Kvamme, Kenneth L.

2006 There and Back Again: Revisiting Archaeological Locational Modeling. In GIS and Archaeological Site Location Modeling. Mark W. Mehrer and Konnie L. Wescott, eds. p 3-38. Boca Raton, Florida: Taylor and Francis.

1980 Predictive model of site location in the Glenwood Springs resource area, in A Class II Cultural Resource Inventory of the Bureau of Land management's Glenwood Springs Resource Area, Nickens, P.R., Ed., U.S. Department of the Interior, Bureau of Land Management, Grand Junction District, CO, Nickens and Associated, Montrose, CO

Lantz, Brett.

2019 Machine Learning with R: Expert Techniques for Predictive Modeling, third edition. Packt Publishing Ltd. Birmingham, UK.

Madry, Scott with Cole, Matthew; Gould, Steve; Resnick, Ben; Seibel, Scott; and Wilkerson, Matt.

2006 A GIS-Based Archaeological Predictive Model and Decision Support System for the North Carolina Department of Transportation. In GIS and Archaeological Site Location Modeling. Mark W. Mehrer and Konnie L. Wescott, eds. p 317-334. Boca Raton, Florida: Taylor and Francis

Meyer, J.G., Urbanowicz, R.J., Martin, P.C.N. et al.

2023 ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 20. https://doi.org/10.1186/s13040-023-00339-9

National Centers for Environmental Information

Climate at a Glance divisional Mapping. Divisional Mapping of North Carolina Divisional Average Temperature 1901-2000 Mean, National Centers for Environmental Information. https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/divisional/mapping/31/tavg/202208/1/mean.

National Hydrography

National Hydrography Dataset. Usgu.gov, https://www.usgs.gov/national-hydrography/national-hydrography-dataset. Accessed September 19, 2024.

National Geographic

Fall Line. Encyclopedic Entry, https://education.nationalgeographic.org/resource/fall-line, accessed December 9 2022.

National Weather Service.

2022 Simplified Hydrologic Cycle. National Weather Service, https://www.weather.gov/mrx/hydro_cycle#:~:text=Water%20infiltrates%20the%20soil%20by,called%20the%20%22water%20table%22., accessed October 25, 2022.

Neumann, Thomas W., Robert M. Sanford, and Karen G. Harry.

2010 Cultural Resources Archaeology: An Introduction. P 93-108, 135-137, 175-176. AltaMira Press, Plymoth, United Kingdom.

North Carolina Office of State Archaeology

2023 Archeological Investigation Standards and Guidelines: For background research, field methodologies, technical reports, and curation. https://archaeology.ncdcr.gov/osa-guidelines/open. Accessed December 2024.

Oguz, F.E., Ekersular, M.N., Sunnetci, K.M. et al.

2023 Can Chat GPT be Utilized in Scientific and Undergraduate Studies?. *Ann Biomed Eng*. https://doi.org/10.1007/s10439-023-03333-8

PA State Historic Preservation Office

2021 Guidelines for Archaeological Investigations in Pennsylvania. PA State Historic Preservation Office (PA SHPO). https://www.pa.gov/content/dam/copapwp-pagov/en/phmc/documents/preservation/about/documents/Guidelines%20for%20Archaeological%20Investigations.pdf . Updated October 2021.

Phelps, David Sutton.

1983 Archaeology of the North Carolina Coast and Coastal Plain: Problems and Hypotheses. In *The Prehistory of North Carolina: An Archaeological Symposium,* edited by Mark A. Mathis and Jeffrey J. Crow, pp. 1-49. Division of Archives and History, Raleigh.

Phillips, Steven

2021 Fitting ‘Maxent’ Species Distribution Models with ‘glmnet’ (R package version 0.1.4). https://cran.r-project.org/web/packages/maxnet/maxnet.pdf

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M.

2011 “pROC: an open-source package for R and S+ to analyze and compare ROC curves.” BMC Bioinformatics, 12, 77.

R-Project

What is R. Introduction to R, https://www.r-project.org/about.html, accessed November 22 2023.

Schleier Jonathan.

2010 GIS Base Archaeological Site Location Modeling in Pitt County, North Carolina. East Carolina Master's Thesis. https://thescholarship.ecu.edu/bitstream/handle/10342/3650/Schleier_ecu_0600M_10333.pdf?sequence=1&isAllowed=n

Shahbaba, Babak

2011 *Biostatistics with R: an Introduction to Statistics through Biological Data.* 1. Aufl. Ed. New York: Springer. doi: 10.1007/978-1-4614-1302-8.

Society for American Archaeology

Ethics in Professional Archaeolog. Career & Practice, https://www.saa.org/career-practice/ethics-in-professional-archaeology, accessed April 13, 2023.

Soil Survey Staff,

Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at https://websoilsurvey.nrcs.usda.gov/. Accessed December, 2024.

The R Foundation

What is R?. r-project.org, https://www.r-project.org/about.html. Accessed April 13, 2023.

Wachtel, Ido, Royi Zidon, Shimon Garti, Gideon Shelach-Lavi.

2018 Predictive Modeling for Archaeological Site Locations: Comparing Logistical Regression and Maximum Entropy in North Israel and North-east China. Journal of Archaeological Science, volume 92 (28-32) https://www.sciencedirect.com/science/article/pii/S030544031830029.

Ward, H. Trawick, R.P. Stephen Davis Jr.

1999 *Time Before History: The Archaeology of North Carolina*. The University of North Carolina Press, Chapel Hill and London.

Wetmore, Ruth Y.

1975 *First on the Land: The North Carolina Indians*. John F. Blair, Winston-Salem.

Wheatley, David, Mark Gillings

2002 *Spatial Technology and Archaeology: The archaeological applications of GIS.* Routledge: Taylor & Francis group.

Wiley, G.R.

1953 Prehistoric Settlement Patterns in the Viru Valley, Peru. Bureau of Merican Ethnology Bulletin 155. Washington D.C.

Yaworsky PM, Vernon KB, Spangler JD, Brewer SC, Codding BF

2020 Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument. PLOS ONE 15(10): e0239424. https://doi.org/10.1371/journal.pone.0239424