ARCHAEOLOGICAL PREDICTIVE MODEL OF SOUTHWESTERN KANSAS

BY

Joshua Stewart Campbell

Submitted to the Department of Geography
and the Faculty of the Graduate School of the University of Kansas
In partial fulfillment of the requirements for the degree of
Master's of Arts

____
Chair

Committee members ______________________

____

____

Date defended:______________

The Thesis Committee for Joshua S. Campbell certifies
that this is the approved version of the following thesis:

ARCHAEOLOGICAL PREDICTIVE MODEL OF SOUTHWESTERN KANSAS

Committee:

Chairperson ______________________________

____

____

____

Date approved: ______________________

Abstract

Knowledge on the archaeological condition of southwestern Kansas is anomalously low, therefore a high-resolution archaeological predictive model has been constructed for the High Plains region of southwestern Kansas. Using quantitative data about the environment as independent variables, the model was constructed using a combination of Geographic Information Systems (GIS) and statistical software. The location of sites was quantitatively related to the environment through a binary logistic regression analysis. The derived regression equation was used to create a unique probability score for each of the 20 million land parcels in the study area. Analysis indicates the model offers a significant increase (30%) over a random classification. 85% of known site locations and 60% of known non-site locations are accurately predicted. In total, the area predicted as site-present comprises 41% of the total study area; within which, the chances of finding a site are 2.15 times as likely as random.

Acknowledgements

Producing this thesis has been a long and arduous process, one that I would not have completed were it not for the help of many individuals. First I would like to thank my advisor, William C. Johnson, who supported my various endeavors and taught me about research, teaching and life in an academic department. Second, the other members of my committee, Brad Logan, Jerome Dobson and Xingong Li, made substantial contributions to my education and subsequently the thesis. Steve Egbert also provided much needed support in the quest for completion. Both the Department of Geography and the Kansas Applied Remote Sensing Program, University of Kansas, provided material support. On a personal level, the support of many friends and family deserves recognition; this group includes (but is in no way limited to) Grandma Maxine (my hero), Dad, Mom and Steve, Matt D, Erich, Rog and Cara, Kasi, Susan, Alex, and the various coffee shops of Lawrence, Kansas.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 1

Introduction

Prior to widespread European settlement of North America, beginning in the 1850s with the passage of the Homestead Act (Scott 1998), the Great Plains (Figure 1.1) were inhabited by indigenous peoples. Archaeological sites, representing a broad range of time, cultures, and activities, have been found throughout the region (Wood 1998). Aboriginal lifeways on the Great Plains consisted primarily of nomadic hunter-gather adaptive strategies for the majority of human occupation, about 11,600 radiocarbon years BP (Holliday 2000). While the exact nature of the hunting and foraging activities of the earliest people in the New World is debatable, a broad range of consensus exists for highly adaptable foraging strategies utilizing an array of available resources.

Regional variations in climatic conditions and landscapes affected adaptive strategies and settlement patterns within the Great Plains. While the development of agriculture (c. 1000 BP) and subsequent cultural evolution to a less mobile, horticulturalist lifeway significantly altered settlement patterns in many areas of the Great Plains (Hofman, Logan, and Adair 1996), in the arid western areas of the Great Plains agricultural practices were not widely adopted and the hunter-gather adaptive strategy remained the primary lifeway.

The level of knowledge about pre-contact indigenous populations varies across the Great Plains. In some locations, primarily in the eastern Plains, European explorers documented large villages, and archaeological excavations have resulted in copious amounts of material culture. The western Plains, or High Plains, are not as well understood. Due to harsh environmental conditions, the majority of archaeological finds in the western Plains represent activities associated with a hunter-gather adaptive strategy based on bison procurement, not intensive agriculture.

Significant numbers of archaeological sites have been found in the High Plains regions of Oklahoma and Texas, yet the same is not true for the High Plains of Kansas (Oklahoma Archeological Survey 2002; Kansas State Historical Society 2002). In Kansas, the density of recorded archaeological sites decreases dramatically from east to west across the state. Approximately 10% of the 12,000 total reported sites are located in the western third of the state, which is coincident with the High Plains physiographic region (Figure 1.2). The distribution of reported sites within the Kansas High Plains is also skewed, with a particularly low number of reported sites in southwest Kansas (Figure 1.3). Eight of nine counties located in the southwestern corner of the state contain less than twenty recorded sites (Kansas State Historical Society 2002). This lack of reported sites seems anomalous considering that several sites of Paleoindian significance are found throughout the adjacent High Plains regions of Colorado, New Mexico, Oklahoma, and Texas and that relatively large numbers of late prehistoric sites are reported from counties in Oklahoma directly surrounding southwestern Kansas (Hofman and Graham 1998; Oklahoma

Archeological Survey 2002; Brosowske and Bement 1998; Bartlett, Bement, and Brooks 1993). The state-line unconformity leads one to question why so few sites are reported in southwestern Kansas.

Recognizing that a large number of reported archaeological sites exist elsewhere on the High Plains, is the low number of reported sites in southwest Kansas due to an actual lack of sites or a lack of formal exploration and reporting? Investigations of archaeological collections from local collectors indicate a significant amount of cultural material has been found in southwestern Kansas (White 2001; Burns 2001). Therefore, the low number of reported sites is most likely due to a dearth of widespread archaeological surveys of the area. Morton County, Kansas is the one exception to the low number of reported sites in the study area (Figure 1.4). Morton County has been extensively surveyed as a result of previous archaeological research and provides a unique opportunity to analyze the spatial pattern of a significant number (>200) of archaeological sites (Brown 1977).

Considering the nature of the archaeological record in the area, some interesting questions arise about the utility of the information recorded in the Kansas State Historical Society registry of recorded archaeological sites. Specifically, in terms of Kansas archaeology, is it possible to use the Morton County data, in combination with the other limited data from southwest Kansas, to make inferences about the landscape choices of native hunter-gather peoples? Can the relationship between the location of known archaeological sites and the environment be modeled using quantitative methods? And ultimately the question becomes, is it possible to use the information about land use patterns derived from a known set of sites to find additional, currently unknown, archaeological sites?

This report attempts to answer those questions through the development of an archaeological predictive model for the High Plains region of southwest Kansas (Figure 1.4). Using Geographic Information Systems (GIS) and statistical software, a probability model has been constructed that empirically relates the presence or absence of archaeological material with nine selected environmental characteristics. The model output identifies areas of the landscape with a set of environmental conditions favorable for finding cultural material. A 'probability surface' was generated in which each of the more than 20 million land parcels in the study area was assigned a probability score for containing cultural material. Each individual probability score was derived from the unique environmental characteristics at each land parcel. Model evaluation was conducted using a set of archaeological sites withheld from model development. Using this method, the power of modern computers and software were used to generate a logical and repeatable inductive quantitative model predicting high-probability areas for finding archaeological site locations.

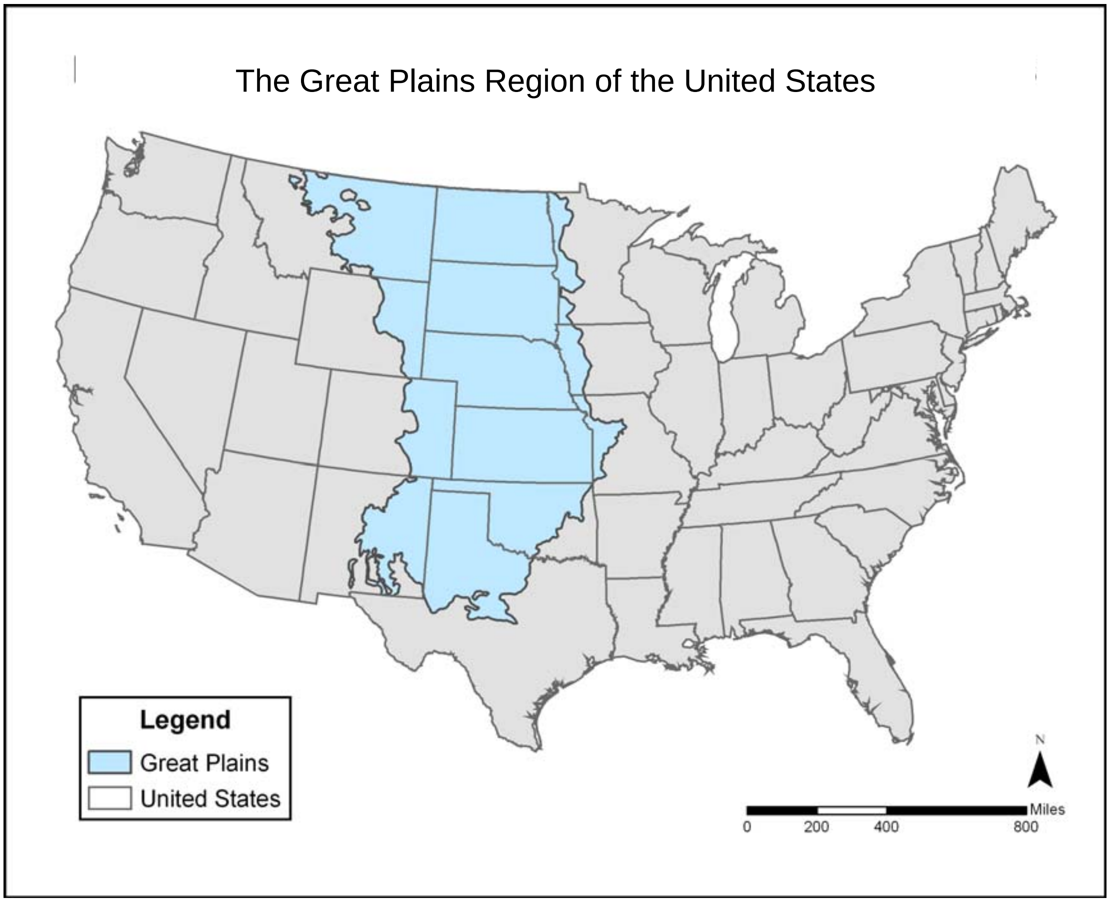


Figure 1.1: Geographic extent of the Great Plains physiographic region in the United States. Extent derived from the Major Land Resource Area (MLRA) classification of the Natural Resource Conservation Service, United States Department of Agriculture.

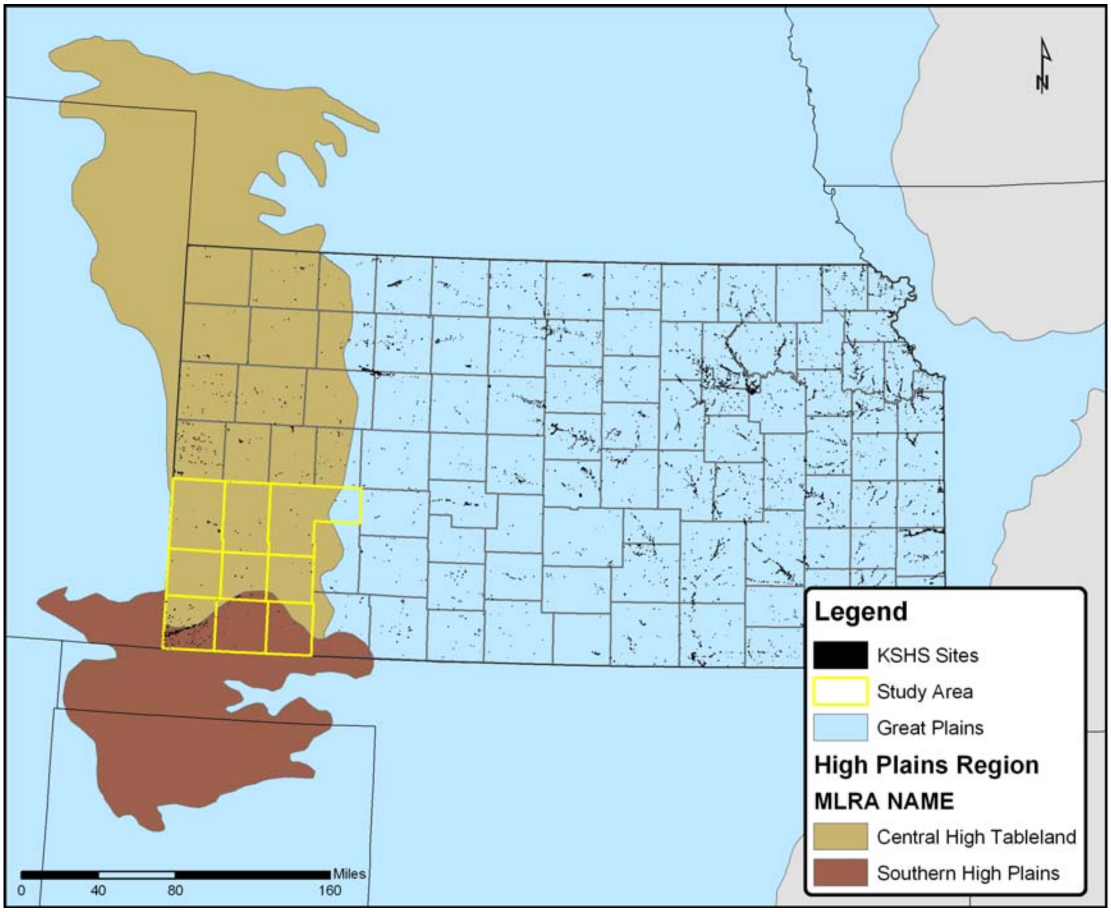


Figure 1.2: Distribution of archaeological sites in Kansas, location of the study area (yellow), and the Major Land Resource Area (MLRA) designations for the study area.

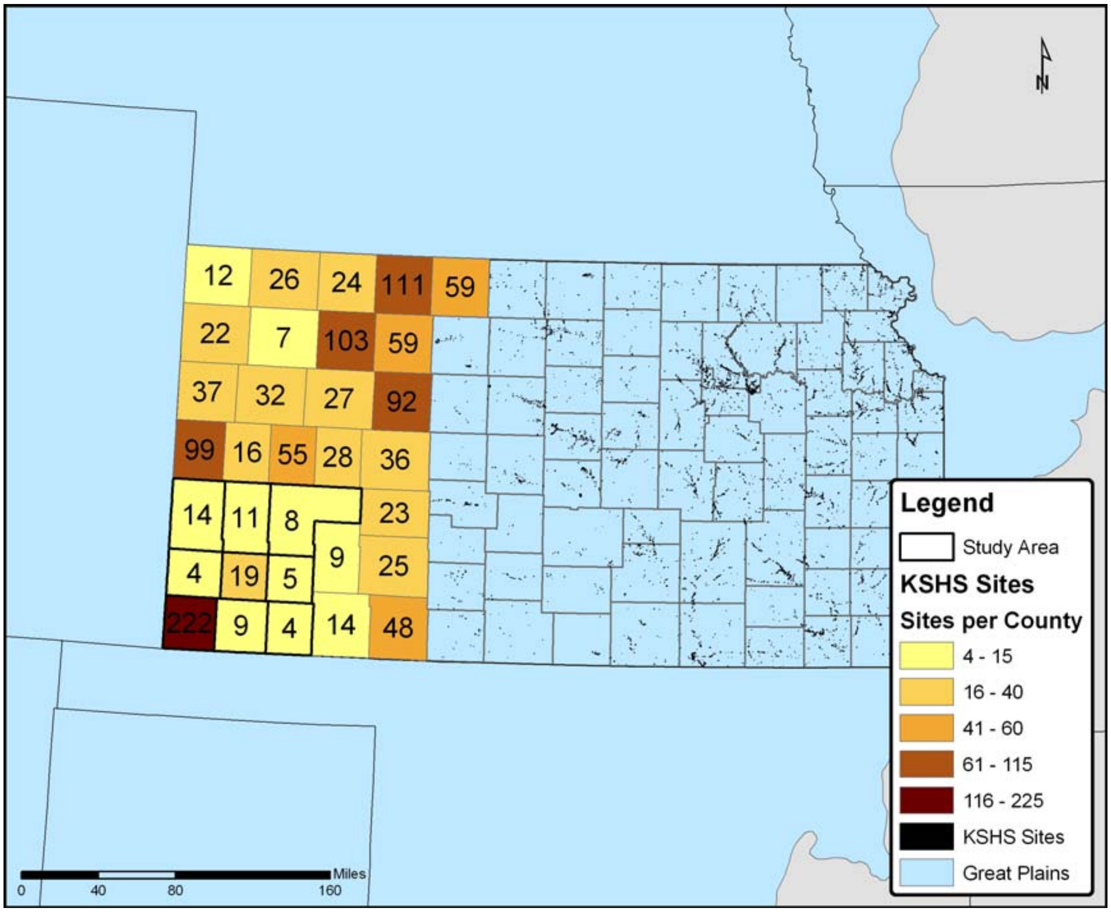


Figure 1.3: Frequency of recorded archaeological sites in western Kansas, by county, within the study area.

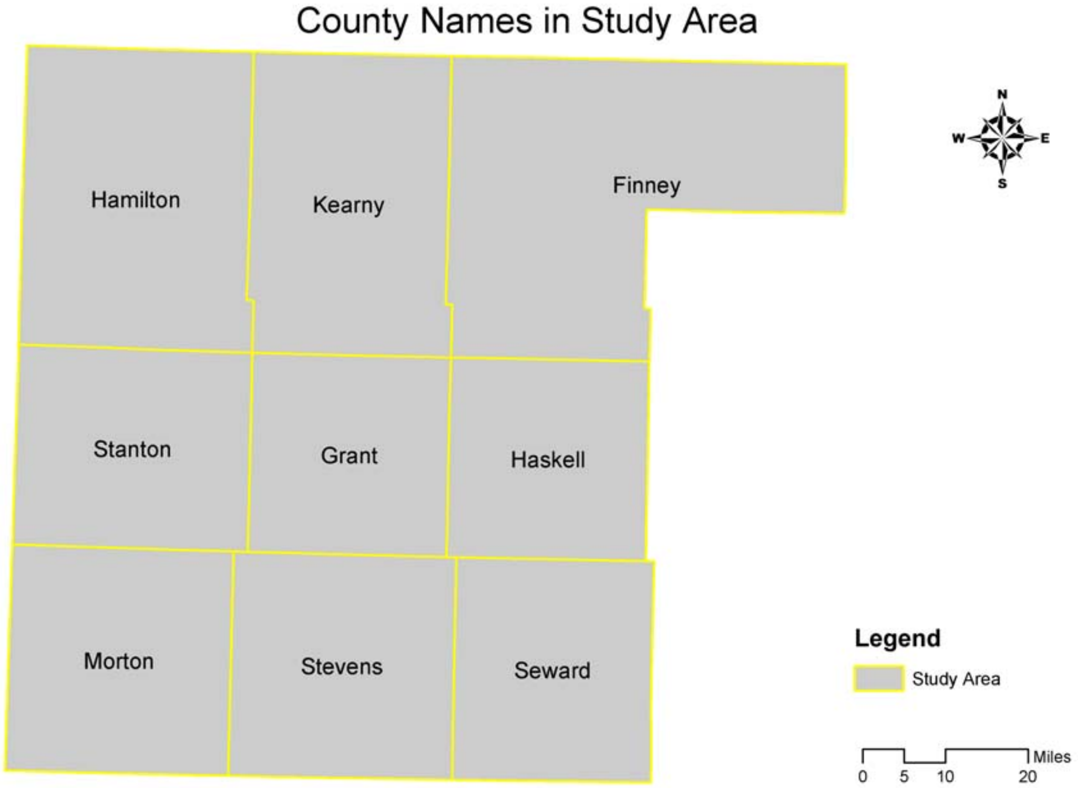


Figure 1.4: The nine counties within the study area, with Morton County being located in the extreme southwest of the study area.

Chapter 2

Archaeological Predictive Modeling Review

This section provides an overview of the theoretical background and range of applications of archaeological predictive modeling. An archaeological predictive model (APM) can be simply defined as a tool that indicates the likelihood of cultural material being present at a location (Gibbon 2000; Warren and Asch 2000). APMs attempt to quantify the spatial pattern inherent to a sample of archaeological site locations with respect to a set of non-archaeological input variables (using any number of pattern recognition methods) and project the abstracted pattern to a larger area (Kvamme 1992).

The theoretical basis of predictive modeling relies on human settlement behavior being non-random, and that the location choices of humans are strongly influenced by the distribution of resources within a certain environment. Therefore, in terms of hunter-gather archaeology, the spatial pattern of archaeological materials on the landscape represents the remnants of an intentional strategy to exploit landscape resources. Pursuit of resources results in a relationship between activity locations and the distribution of certain environmental resources. Predictive models assume the environmental factors that influenced settlement choices are accurately represented in modern maps of environmental resources (Warren and Asch 2000); therefore information extracted from modern maps can be used to explain the distribution of activity locations.

A more specific explanation of archaeological predictive models is offered by Kvamme (1990:261), who defines a predictive model as "an *assignment procedure* that correctly indicates an *archaeological event outcome* at a *land parcel location* with greater probability than that attributable to chance." The assignment procedure, or decision rule, is a set of criteria that classify a land parcel into an archaeological event class on the basis of some non-archaeological input. For many APMs, the decision rule uses environmental information about a land parcel as input variables. Output of a decision rule is the classification of the land parcel to an archaeological event class (Kvamme 1990). Environmental-based APMs determine the probability that a site occurs at a given location by measuring an appropriate set of environmental variables (Warren and Asch 2000). Each of these three components is discussed in greater detail below.

Predictive models can be divided into two main groups based upon the type of decision rule. *Inductive predictive models* utilize statistical techniques to determine the quantitative relationship between site locations and the environment. In contrast, *deductive predictive models* use intuition or deductive reasoning to model the relationship between archaeological sites and the landscape. A professional archaeologist with significant experience in a particular region could construct a deductive model based on some set of characteristics believed to influence the distribution of sites. Hudak et. al. (2000) compared the accuracy of deductive and inductive models created for an area of Minnesota and reported the relative strengths and weaknesses of both types of models. While the deductive model, created by a trained archaeologist familiar with the area, performed well when compared with the earliest modeling efforts (Phase 1), the most advanced inductive models (Phase 3) were more accurate by a statistically significant margin. The focus of this report is on inductive modeling methods.

A survey of the predictive modeling literature indicates the practice of inductive archaeological predictive modeling was well established by the mid-1980s (Judge and Sebastian 1988; Carr 1985). The primary motivation behind the development of such models in the United States originated with federal land management agencies, such as the U.S. Army, Bureau of Land Management and the U.S. Forest Service. The increasing availability of powerful computer hardware and Geographic Information Systems (GIS) software in the early 1980s, combined with legislation dictating the management and protection of cultural resources on federal lands (National Historic Preservation Act of 1966), provided the means and incentive for the development of computationally intensive archaeological predictive models. Inductive models had been previously constructed, but the large number of computations and map data extractions made their implementation difficult (Pilgram 1987); GIS and digital spatial data provided the first digital tools for the construction and development of large inductive predictive models (Kvamme and Kohler 1988).

A seminal article written by Kenneth Kvamme (1990), entitled *The Fundamental Principles and Practice of Predictive Archaeological Modeling*, provides the theoretical foundation of archaeological predictive modeling and represents an attempt to place the varied practices of inductive and deductive predictive modeling into a common conceptual framework. This article follows on Kvamme's early work on the Pińon Canyon models, conducted for the US Army in 1983, and a chapter entitled "Development and Testing of Quantitative Models" in Judge and Sebastian (1988) *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*. Methodologies for inductive predictive modeling and accuracy assessment presented in Kvamme (1990) have been widely adopted by various researchers (Dalla Bona 2000; Duncan and Beckman 2000; Hudak et al. 2000; Krist 2001; Lock and Stancic 1995; Lock 2000; Premo 2001; Westcott and Brandon 2000; Wheatley and Gillings 2002; Warren and Asch 2000; Wescott and Kuiper 2000). In the 1990 article, Kvamme articulates how to develop, and more important, how to evaluate the effectiveness of an APM. Regardless of whether an inductive or deductive approach is used, a successful APM minimizes the classification error (sites versus non-sites) well enough that it represents a significant gain in accuracy over random chance models (Warren and Asch 2000). Kvamme's approach forms the methodological basis for this current project.

Use of predictive modeling has increased as GIS software and computer hardware costs have dropped (Allen, Green, and Zubrow 1990; Lock 2000; Lock and Stancic

1995; Westcott and Brandon 2000; Wheatley and Gillings 2002). Another factor contributing to the recent increase in the development of predictive models is the widespread availability of georeferenced, digital geographic data, *e.g.* elevation, soils, hydrology. The utility of GIS for archaeological applications was recognized early on by the archaeology community, as evidenced by the establishment of a GIS section at the annual "Computer Application and Quantitative Methods in Archaeology" conferences in 1990. Growth in archaeological applications of GIS has been driven by the advancement of geographic information science and the macro trends in information technology of increasing processing power and decreasing costs.

Predictive modeling was developed to increase the financial efficiency of surveys for cultural resource management (CRM). CRM represents an alternative approach to archaeological exploration than that offered by academic research. CRM is based upon quantifying the extent of cultural resources and managing/protecting those resources (Verhagen 2000; Lang 2000). Ground surveys are widely used during CRM projects; these surveys are expensive in both work time and travel costs. Optimally, archaeological surveys would minimize the amount of energy spent in terms of money and work hours for field survey and maximize the return, namely the amount of archaeological data recovered. Because the resources devoted to survey usually are limited, a method of relating the presence of archaeological materials with certain landscape features would increase the efficiency of CRM survey design (Hudak et al. 2000).

GIS-based APMs are essentially macro-scale landscape screening tools. The screening component of APMs occurs when the empirical relationship extracted from the sample data is projected onto areas not surveyed for archaeological sites. GIS allows the compilation of datasets covering very large tracts of land; using an APM, these large unsurveyed areas can be 'screened' for the potential of containing sites. In terms of CRM, the utility of predictive models can be measured from three distinct perspectives. First, they provide archaeologists a reliable picture of the potential distribution of sites for optimizing field surveys. Second, land managers can use these models to make decisions regarding the preservation of cultural resources. Third, they provide land developers with the ability to plan construction projects in areas where cultural resources are expected to be low. In each case having a reliable predictive model is economically prudent (Warren and Asch 2000).

The most ambitious predictive modeling effort undertaken to date was created for the State of Minnesota and is known as the Mn/Model. Funded by the Federal Highway Administration, the Mn/Model is a collection of models generated for each physiographic province of the state; therefore the final model consists of 24 different regional predictive models. Three distinct modeling phases are detailed in the Mn/Model final report. Overall, the large geographic extent of the project distinguishes it from other applications of predictive modeling. Two reasons for its success were first, the development of an integrated professional staff of archaeologists, geomorphologists, and GIS specialists and second, the establishment of procedures, workflows, and accuracy criteria (Hudak *et al*., 2000). Developing the

Mn/Model required a large operating budget (approximately 4.5 million dollars) and a large staff. In 1997, the year before the Mn/Model project began, the State of Minnesota spent 1.7 million dollars on site mitigation projects per construction season. Average savings since the completion of the model are estimated at 3 million dollars per year. Using these estimates, the modeling effort more than paid for itself in less than 2 years. The States of North Carolina and Vermont are currently in the process of developing statewide APMs, again under the direction of their respective Departments of Transportation.

Fundamental Components of Predictive Models

As stated previously, Kvamme's (1990) definition emphasizes the importance of predicting an "archaeological event at a land parcel with greater probability than that attributable to chance". From this definition, three fundamental components of an APM are recognized: the unit of analysis as a land parcel, the development of an assignment procedure, and the application of the assignment procedure to each land parcel to assign the parcel to an archaeological event class. In the following sections, each of these three fundamental components of archaeological predictive models are discussed, as well as the process of variable selection and several critiques of the methodology.

Unit of Investigation

The fundamental component of any archaeological predictive model is the unit of investigation. Typically, in archaeological studies the analysis unit is the archaeological site. In the case of archaeological predictive modeling, however, the unit of investigation is the individual parcel of land (Kvamme 1988). In an APM, the entire study area is divided into discrete parcels of uniform size. Dividing the landscape into a series of contiguous land parcels works well with the use of GIS, as the single land parcel forms the standard grid cell used in raster data analysis. The assignment procedure component of the APM is then applied to each grid cell / land parcel.

Determining the appropriate pixel resolution / land parcel size involves consideration of the modeling goals and the available geographic data. In terms of modeling goals, most common is the prediction of site locations. Other goals involve the prediction of the number of sites within a land parcel (typically used when the study area is very large and the available geographic data is limited) or the number of artifacts within a parcel (usually implemented at the individual site-level). Thus, selection of the land parcel size has implications for the model output. A large parcel size (>1km) may be useful for predicting the number of sites within a given parcel, however, this parcel size may be too coarse for predicting specific site locations. Predicting artifact density is typically conducted within small geographic areas and utilizes a small land parcel, *e.g.* excavation unit (1m^2); this parcel size is too small for predicting site locations because the resolution of the input environmental data is much coarser than the excavation unit.

The optimal parcel size for predicting site locations captures the variability of the landscape that influenced cultural behavior but is not a finer scale than the available environmental data (Hudak et al., 2000). Considerations of the available environmental datasets are important because spatial datasets are collected with a specific margin of error and consequently have limits to the positional accuracy of the data (Clarke et al., 2002). Use of a land parcel size that is at a finer resolution than the mapping scale of the geographic data risks the introduction of error or false precision into the model. Most APMs created for the United States in the last several years use a $30m^2$ land parcel. This choice stems from the widespread availability of 1:24,000 scale Digital Elevation Models, distributed by the United States Geological Survey, that have a $30m^2$ pixel size. Kvamme (1992) reports that a $50m^2$ land parcel size would be considered a moderate to high-resolution model for predicting site locations. A previous model created by the author for Fort Hood, Texas utilized a $5m^2$ pixel. Utilizing this small pixel size was possible because of the high-resolution terrain and hydrologic mapping data available for the base (Campbell and Johnson 2004).

Archaeological Events

Output of an archaeological predictive model is the assignment of a land parcel to an archaeological event class. Archaeological event classes must be defined prior to model construction. The simplest set of archaeological events involves classifying a parcel into either a *site-present* or *site-absent* class. Archaeological event classes can also be structured to predict the type of site present at a location, the number of sites within a parcel, or the density of artifacts within a parcel. Regardless of the modeling goals, the set of potential event classes must be mutually exclusive and exhaustive, meaning a parcel must be assigned to only one of the event classes and all parcels must be classified (Kvamme, 1990).

Using notation derived from Kvamme (1990), the following section describes the potential event classes used in a typical site prediction model. For each land parcel used to construct the model, two potential archaeological events representative of the true condition of the land parcel are possible:

S = {site-present}

or

S’ = {site-absent}

Output of the model is the assignment of every land parcel into one of two potential archaeological event classes:

M = {model predicts site-present}

or

M’ = {model predicts site-absent}

The difference between these two sets of event classes is crucial for interpreting model results. Any single land parcel can be classified according to its condition in reality (S or S') and by its condition predicted by the model (M or M'). Because no model makes perfect prediction, the true condition and the model prediction of a land parcel may not agree. Comparing the relative values of S, S', M, and M' provides a quantitative method for evaluating model performance. This notation is used throughout this report.

Predictive Models as Decision Rules

An archaeological predictive model is essentially a decision rule conditional on other, non-archaeological features of a location (Kvamme, 1990:261). Decision rules can be generated using techniques ranging from an inductive analysis, using statistical techniques to derive an equation from empirical patterns in sample data, to a deductive analysis in which a trained archaeologist creates decision rules based on previous knowledge of cultural patterns. It is reasonable to assume that indigenous people chose site locations based upon a simultaneous consideration of multiple environmental criteria. The critical question when constructing an archaeological model is the relative weights to associate with each non-archaeological variable. A professional archaeologist working within a region will inevitably have a mental conception of where sites occur on the landscape, *e.g.,* 30 meters from stream, or less than 10% slope. However, this information is often geographically localized and may vary between archaeologists. The utility of statistical techniques is the unbiased and independent method in which the effect of a specific variable on site location is derived; in terms of a regression-based predictive model, the importance or weight of each variable is essentially the coefficient applied to each variable. Deductive knowledge is required for the initial variable selection, but the specific weights for each variable are derived from the spatial patterns of the sample data. In this way, a researcher can focus on the selection of appropriate variables and data structures and allow the statistical method to derive the variable weights. Inductive methods are independent of personality and experience, and, because the results are based solely on the input data, results are reproducible. Kvamme (1990), Carr (1985), and Parker (1985) provide a thorough review of various statistical and inductive methods. For multiple reasons, the predominant statistical technique used in inductive archaeological predictive modeling is the logistic regression method. Binary logistic regression is discussed in greater depth in the following sections.

Factors Influencing Variable Selection

A survey of the available literature indicates inductive predictive modeling has been utilized in various geographic and archaeological contexts within North America and Europe (Allen, Green, and Zubrow 1990; Lock 2000; Lock and Stancic 1995; Westcott and Brandon 2000; Wheatley and Gillings 2002). While the statistical techniques used for model construction and testing are independent of geography and culture, the variables used within the models to explain cultural behavior are sensitive to these factors and must be considered prior to model construction. Primary factors to consider in model design are the type and complexity of the economic system inherent to the cultural group under study and the landscape in which the cultural group operated. Selection of relevant variables for model inclusion is dependent upon the mechanisms in which the cultural group under study interacted with the environment. Consider the differences between a nomadic hunter-gather on the Great Plains (Warren 1990b), a sedentary horticulturalist in the Appalachian Mountains (Duncan and Beckman 2000), and a Roman agriculturalist on a Mediterranean island (Stancic and Veljanovski 2000). Clearly the relationship between cultural activity and environment are different in these situations, thereby leading to a different set of relevant environmental variables selected for entry into the model.

Hunter-gather lifeways can be described as following an optimal food procurement strategy in which the culture group extracts a living directly from the environment and patterns its site selection on the basis of minimizing energy output. For hunter-gatherers, fundamental resources relate to the procurement of water, food, and shelter (Bamforth 1988; Butzer 1982; Jochim 1976; Wedel 1963). In contrast to hunter-gatherer groups, the market-driven economic systems of more advanced societies, primarily in Europe with some examples in North America, result in site patterns not entirely based on environmental resources. In these cases, social factors (distance to road, distance to agricultural soil type, viewshed of defensive fortifications) may be important for describing site patterns (Wheatley 1995; Wheatley and Gillings 2002). Appropriate variable selection requires a theoretical understanding of the culture, environment, and time period under analysis. Spurious correlations may occur if inappropriate variables are included in model development; a model may be technically accurate but not have any real archaeological meaning.

Landscape variation also influences cultural behavior, and therefore geographic considerations must be factored into the modeling methodology prior to analysis. It is reasonable to assume the importance of any given variable will change across space. For example, a distance to water variable may be critical in arid regions, but not significant in a tropical climate. One method of dealing with environmental variation is to divide the landscape into distinct physiographic regions and model each region separately (Hudak et al., 2000). Resource distribution within a region will influence site patterns. Therefore, if the distribution of resources or the type of resources change significantly among regions, then a model constructed for one region may not be appropriate for another region. Regional division of the landscape can be based upon any physiographic criteria, so long as the divisions represent significantly different resource zones. It is important to note that if a quantitative method is used in the modeling process, the derived equation should only be implemented within the region it was developed.

Geographic Information Systems

The development and growth in the use of APMs is tied to the development and accessibility of Geographic Information Systems (GIS) (Westcott and Brandon 2000). GIS provide digital methods for the storage, analysis and visualization of spatial data. These methods are essential to the construction, implementation, and testing of archaeological predictive models. The size of datasets and the immense number of calculations required to compute a model requires the power of modern computing hardware and software. Kvamme and Kohler (1988) present a thorough review of the use of GIS in archaeological predictive models; although the software has changed substantially since 1988, this work still provides valuable information in regards to the basic algorithms used in spatial modeling.

Construction

GIS is the central management hub for the compilation of data sources and the extraction of data for model construction. Using GIS, large environmental datasets can be constructed and stored for later use. Many predictor variables are derivatives of primary datasets, *i.e.* slope is derived from a digital elevation model, and GIS provides a toolkit for the creation of new derivative data layers. The numerical data used to create a model are extracted from the GIS database and exported for use in a statistics package. Prior to GIS all measurements were taken manually from maps, effectively prohibiting the size of input data pools and the size of land parcels (Pilgram 1987).

Implementation

When used in conjunction with GIS, predictive models can be thought of as macro-scale landscape screening tools. The prediction, or screening, component of these models occurs as the empirical relationship extracted from the sample data is projected onto areas where the archaeological distribution is not well understood. A quantitative APM is an equation, created using statistical software, which can be applied to any given land parcel in the study area. With GIS and 'Cartographic Modeling' the finished model equation can be applied to every land parcel within a study area in a matter of moments (Tomlin 1990).

By applying the quantitative abstraction to the entire study area, the model selects locations with a set of landscape characteristics similar to those of the input sample of known site locations. Identifying these areas of the landscape should increase the likelihood of finding unknown sites. In terms of CRM or development planning, having information about the potential location of sites can save time and resources (Hudak, 2000). Prior to GIS, projecting the completed model into unsurveyed areas required the manual application of the quantitative model to each selected land parcel. The number of measurements and calculations required to compute the model for any given size required too much work to be effective (Pilgram 1987).

Assessment

Determining the accuracy of a model also utilizes the GIS toolkit, both in terms of visual inspection of the model and analytical evaluation. Visualization capabilities of GIS allow a visual representation of the model as it applies to every land parcel, thereby allowing an analyst to visually interpret the spatial pattern of the model results. While not quantitative in nature, a visual analysis of the probability surface created by applying the model to each land parcel is a valuable investigative tool for exploring the spatial implications of the model. Because the model output for each land parcel is based on multiple input variables, interesting spatial patterns only appear when the output equation is mapped as a continuous surface. Quantitative model evaluations utilize the same data export tools used in model construction. Predicted probabilities for each land parcel in the model-testing group are exported to a statistics program for quantitative assessment and graphic production.

Critiques

Questions relating to the philosophical implications of predictive models have been widely explored in recent works (Church, Brandon, and Burgett 2000; Ebert 2000; Gaffney and Leusen 1995; Kuna 2000; Lang 2000; Verhagen 2000). The fundamental issue relating to the acceptability of archaeological predictive modeling involves the extent to which model outputs are environmentally deterministic. The question becomes: Is it acceptable to predict human behavior using only environmental variables? Two different perspectives have emerged on the issue; these perspectives are closely related with a larger question involving the division between academic archaeological research and CRM.

Opponents of inductive predictive modeling argue that environmentally deterministic models do not offer a substantive explanation into the nature of the archaeological distribution (Ebert 2000; Gaffney and Leusen 1995). By focusing solely on the environmental considerations of site location, the role of culture or human agency is overlooked. The simplistic explanations provided by environmentally deterministic models are therefore not truly ‘archaeology’ because they do not begin with a holistic approach to the study of previous cultures. This view generally corresponds with the perspective of the archaeological research community.

The primary critique on the use of geographic data as predictor variables involves the reliability of available spatial datasets to accurately represent environmental conditions in the past (Ebert 2000; Gaffney and Leusen 1995). Essentially the question becomes, for instance, how relevant is a ‘distance to water’ variable if the river has changed its course through time? In addition, opponents argue that the location of the river may have had no impact on site location and that some cultural or other immeasurable factor influenced site selection. Opponents point to GIS as contributing to the expansion of a theoretically bankrupt methodology (Ebert 2000). The creation of GIS datasets is not trivial in terms of time and effort, and, as a result, most predictive models utilize available environmental data, *e.g.,* soils, hydrology, elevation and geology. Opponents argue that inappropriate environmental variables are used for models simply because they are available.

Those in favor of environmentally based predictive models point to success of models in CRM, which is concerned with the protection of cultural resources, not necessarily the same goals as academic research in the holistic explanation of culture (Gaffney and Leusen 1995; Lang 2000). In a CRM context, the use of environmental data to describe the potential locations of cultural material is hard to resist, particularly in light of the fragmentary nature of the known archaeological record. Artifacts represent the material residue of some activity in the past that has survived to be discovered in the modern period. Often the only materials to survive are lithic tools, which tended to evolve slowly over time and are difficult to assign to a particular 'culture'. The lack of resolution in the archaeological record concerning the age and function of materials makes it difficult, if not impossible, to measure the subtle influence of cultural, *i.e.,* non-environmental, forces on the creation and deposition of material. Therefore the detection of additional archaeological resources, by any means, is useful.

In contrast, many archaeological sites demonstrate repeat habitation, indicating the environmental resources of a location are found desirable by different cultures throughout time. If environmental resources are consistently found desirable, and those resources change slowly through time, then searching for unknown cultural materials on the basis of environmental conditions can be justified, particularly if the goal is cultural resource management. Well-constructed archaeological models accurately predict 70 - 85% of known archaeological sites. Repeated credibility of such accuracies indicates the relevance of predictive modeling as an investigative tool (Gaffney and Leusen 1995; Hudak et al. 2000). The Mn/Model in particular has illustrated the success with which predictive models can be incorporated into CRM. Additional statewide probability models, also known as archaeological sensitivity models, have been produced in North Carolina and most recently Vermont (ArcNews 2006).

The utility of geographic variables for predicting human settlement patterns is not confined to the archaeological record. The LandScan Global Population Project represents the most comprehensive global model of human population distribution patterns available today (Dobson et al. 2000). The best available census counts for a given area, often at the provincial level, were redistributed to 30 x 30 arcsecond grid cells on the basis of distance to roads, slope, land cover, and (formerly) nighttime lights. Considering that modern human populations can be accurately modeled by geographic variables, it is not a giant conceptual leap to assume that cultures more reliant on a direct extraction of resources from the environment would be influenced by similar factors.

Empirical correlations generated by predictive models should be viewed as providing insight into where cultural materials are located, not explicitly defining why the materials are there. The influence of human agency in cultural adaptation cannot be easily integrated into a numerical analysis of site patterns. However, increasing knowledge about the 'why' of past cultures fundamentally requires new archaeological data for analysis. Predictive models are effective tools for locating unknown cultural resources and should not be discarded because they do not offer holistic cultural explanations (Hudak et al. 2000; Warren and Asch 2000). Predictive models will not replace human investigation of the landscape; however the application of a macro-scale landscape screening tool will improve research design, thereby resulting in more efficient archaeological surveys and cultural resource management (Verhagen 2000).

Logistic Regression

The predominant method used in constructing quantitative archaeological predictive models utilizes a logistic regression technique, either binary or multivariate. Binary logistic regression, a type of probability model, is useful when the observed outcome is restricted to two values, which in this case represent the *site-present* {S} and *site-absent* {S'} event classes (Warren 1990a). These events are coded as 1 and 0 respectively, for use in the database. Output of the binary logistic regression represents the probability of the event occurring, expressed as the *Prob(event*) or in this case the probability of a site occurring *Pr(M)*. In ordinary regression, the output value of the equation (Z) can be any value, positive or negative. Because the logistic model output is a probability, the output must be constrained between 0 and 1. Ordinary regression output (Z) must be converted to a probability value constrained between 0 and 1 (Clark and Hosking 1986). The standard linear regression equation can be generically described as:

$$\mathbf{Z = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p}$$

where, Z is the predicted output of the regression equation (dependent variable), B_0 is a constant term, B_p is a coefficient and X_p is an independent variable for every variable in the equation. In order to convert the raw output to a probability of the event occurring, the following equation must be applied where ***e*** is the natural log and (-***Z***) is the ordinary regression output multiplied by -1:

$$\mathbf{Pr(M) = 1 / (1 + \boldsymbol{e}^{-Z})}$$

And conversely, the probability of an event not occurring is expressed as:

$$\mathbf{Pr\ (M') = 1 - Pr\ (M)}$$

Preference for logistic regression is based upon multiple factors. The method is robust with respect to the data normality and equality of variance assumptions required of related techniques, *e.g.,* discriminant functions, and it can also handle nominal, ordinal, ratio, or interval level data (Gibbon 2000; Kvamme 1990; Parker 1885; Warren and Asch 2000). Kenneth Kvamme developed the method for use in archaeology in the early 1980s (Warren 1990a); Kvamme's method of model development and assessment is used for the model described herein.

Chapter 3

Study Area

As stated in the introduction, the study area of this report consists of nine counties located in the High Plains region of southwest Kansas (Figure 1.2 and Figure 1.4). Selection of this area was based on multiple criteria, primarily the nature of geospatial data storage and physiographic homogeneity. Political boundaries (county divisions) are used as the structure of geospatial data storage at the Kansas State Historical Society (KSHS) and the Data Access and Support Center (DASC). Although political boundaries define the exterior of the study area, this approximately 85 mi. x 75 mi. area was selected on the basis of physiographic homogeneity and poorly understood archaeological record.

In terms of archaeological modeling, it is important to consider the size, extent, and landscape variability of the physical area to be modeled. Optimally, a predictive model should be developed for a homogeneous land area. If a significant change in the distribution of natural resources changes within a modeled area, the pattern of landscape utilization also changes. The predictive power of an archaeological model is based upon demarcating the landscape into homogenous tracts in which the available environmental resources, and subsequently the adaptive strategies utilized by humans to extract those environmental resources, are considered similar.

Therefore it is critical that the area selected for model inclusion have a large degree of similarity in the distribution of landscape resources (Hudak et al. 2000).

The definition of physiographic homogeneity can be based on many perspectives and associated classification schemes including geologic, vegetation community, and pedological. From a geological perspective, the area is classified as the High Plains and Arkansas River Lowlands (Kansas Geological Survey 1984); pedologically, the area straddles the Major Land Resource Area (MLRA) boundary between the Southern High Plains and Central High Tablelands (Figure 1.2) (Soil Survey Staff 1981). A.W. Kuchler's map of the "The Potential Natural Vegetation of Kansas" displays the vegetative similarities of the study area; note the nearly continuous distribution of short-grass prairie, with tracts of sandsage prairie along the south side of both the Arkansas and Cimarron Rivers, and floodplain vegetation (Figure 3.1). Ultimately the pedological perspective was chosen as the basis for determining physiographic homogeneity. The MLRA system was chosen because it encapsulates components of geology, climate, vegetation, and landforms into a coherent whole.

Major Land Resource Area (MLRA)

In order to understand the affect of landscape on archaeological predictive modeling, a basic review of the geographic perspective used to characterize the physical environment of the study area is required. Specifically, a pedologic perspective is used to organize the macro-landscape into smaller regions of similar physical characteristics. The Land Resource Region (LRR) and Major Land Resource Area

(MLRA) system developed by the Natural Resource Conservation Service (NRCS) is used (Soil Survey Staff 1981). The NRCS, as the federal agency responsible for the identification, cataloging, and mapping of soils in the United States, has developed a hierarchical system for segmenting the U.S. into regions of homogeneous physical units on the basis of soil properties. Soil formation is a function of several factors operating simultaneously, commonly referred to as CLORPT (CLimate, Organisms, Relief, Parent material, and Time). Each of these landscape characteristics is considered when designating a soil unit; therefore individual soil series represent a unique combination of landscape factors. Soil series create a high-resolution view of the landscape (1:24,000 scale) and reflect the finest level of variability among soil units recognized by the NRCS.

However, many soils in a region share some measure of similarity. As the area of the landscape under analysis increases, it is reasonable to group similar soils into larger physical units. Hierarchical organization of soil units from a high-resolution soil series level into larger macro-areas composed of similar soil series is the basis of the LRR\MLRA system. Organizationally, the system begins at a single geographic location or the pedon level (1:1 scale), and extrapolates out to the Land Resource Region level (1:7,500,000 scale). Every soil pedon can be grouped into some aggregation at the component (1:10,000), SSURGO (1:24,000), STATSGO (1:250,000), Land Resource Unit (LRU) (1:1,000,000), MLRA (1:3,500,000), and finally LRR level, depending upon the desired resolution of landscape information (Figure 3.2).

For the purposes of this research, the study area is considered the Southern High Plains (Figure 1.2). This decision is based upon the "Major Land Resource Area" divisions developed by the Natural Resource Conservation Service. Although the MLRA division between the Southern High Plains and Central High Tablelands is the Cimarron River, it is a transitional area between both of the MLRA divisions and displays more internal similarity than the classification requirements of either the Central High Tablelands or the Southern High Plains. Additional evidence supporting the transitional concept for the study area comes from the archaeological literature; this area has been classified as both the Southern High Plains (Wood 1998) and the Central High Plains (Hofman 1996).

Physiography of the Study Area

The study area can be characterized by the broad expanses of short-grass prairie uplands separated by entrenched fluvial systems. Relief on the uplands tends to be low, although contrary to popular belief, not entirely flat. Overall the landscape is highest in the west and gently slopes eastward; the difference between the highest and lowest elevation values in the study area is 1,430 feet. In southwestern Kansas, the upland plains are separated by two large river systems, the Arkansas and Cimarron (Figure 3.3). These two rivers are significantly different in terms of flow regime and channel morphology. Both provide excellent locations to view large expanses of the surrounding terrain, such as the Point of Rocks area in Morton County (Figure 3.4). Surface water is scarce in the study area; outside of the two large river systems, the majority of the stream networks exhibit intermittent surface flow. Stream networks dissect the uplands and the large river valleys are entrenched more than 50 feet below the surrounding surface (Figure 3.5). Playa lakes, with occasional associated lunettes, are a common feature of the uplands. From an archaeological perspective, playa lakes in the Southern High Plains of Texas were an important resource. Excavations at Lubbock Lake and the Miami site indicate the archaeological significance of these features (Holliday 1997). Johnson and Campbell (2004) created a playa database for the state of Kansas that was adapted for this project (Figure 3.6).

Dune fields are another important geomorphic feature of the study area; large dune fields formed from aeolian erosion of sediment sources in the river valleys. The primary deposition on the south side of the valleys is a result of the predominant northwestern late-Pleistocene wind direction (Johnson and Park 1996). Some remobilization of dune sand has occurred during the Holocene. Portions of the Cimarron dune field have remobilized and sand has migrated to the north side of the river valley. Figure 3.7 displays a landform map generated for the study area. Landforms were generalized from the SSURGO data into six landform categories using the MLRA concepts discussed above; the dataset is discussed further in the following section on Environmental data.

High-quality lithic resources within the study area are few, although some utilization of local quarries and other local materials have been identified in the archaeological database (Kansas State Historical Society 2002). The Alibates quarry is less than 100 miles south of the study area, and Alibates chert artifacts have been found in the study area (Brown 1977). Locally derived lithic sources include quartzites and orthoquartzites from the Ogallala and Dakota formations and various cherts and petrified wood from alluvial deposits of the Arkansas and Cimarron rivers (Brosowske and Bement 1998). These resources indicate that both locally derived and imported lithic materials are expected within the study area.

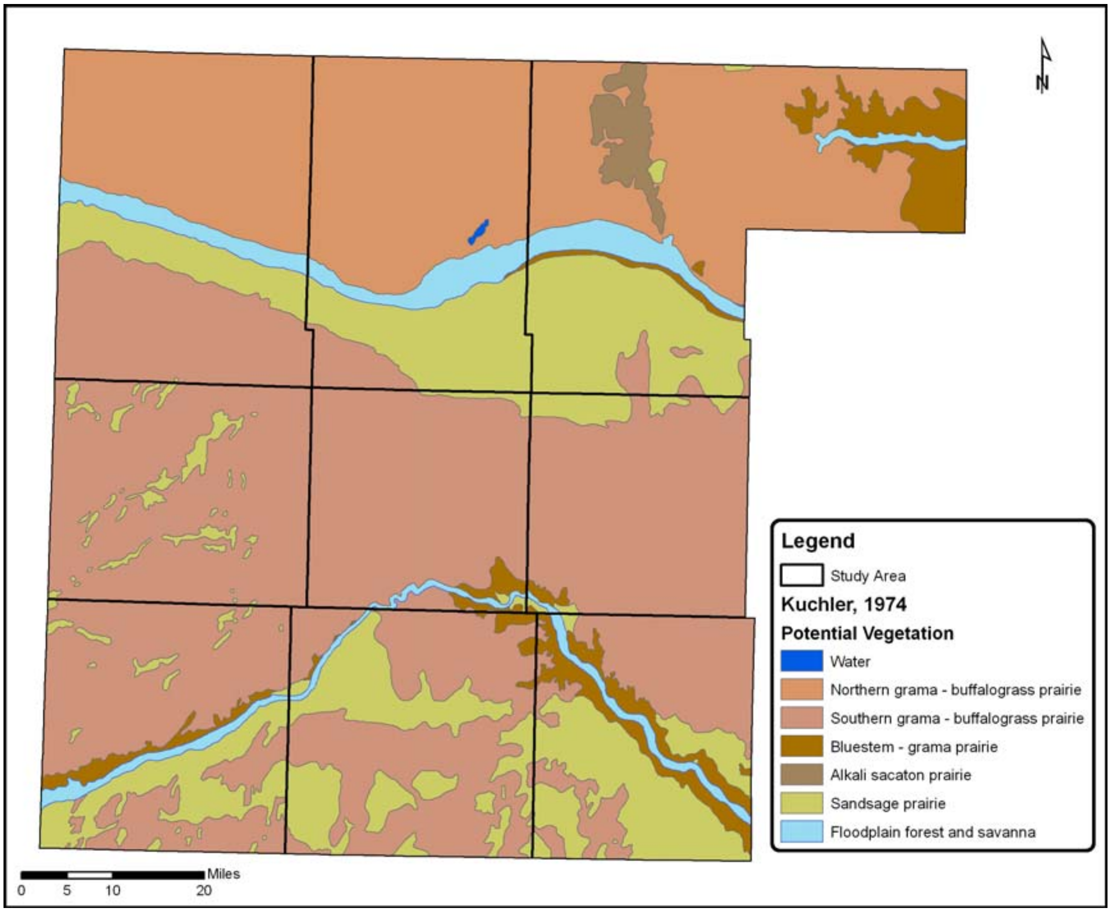


Figure 3.1: Distribution of Potential Natural Vegetation of Kansas. Similarity of potential vegetation indicates that the study area is physiographically homogeneous.

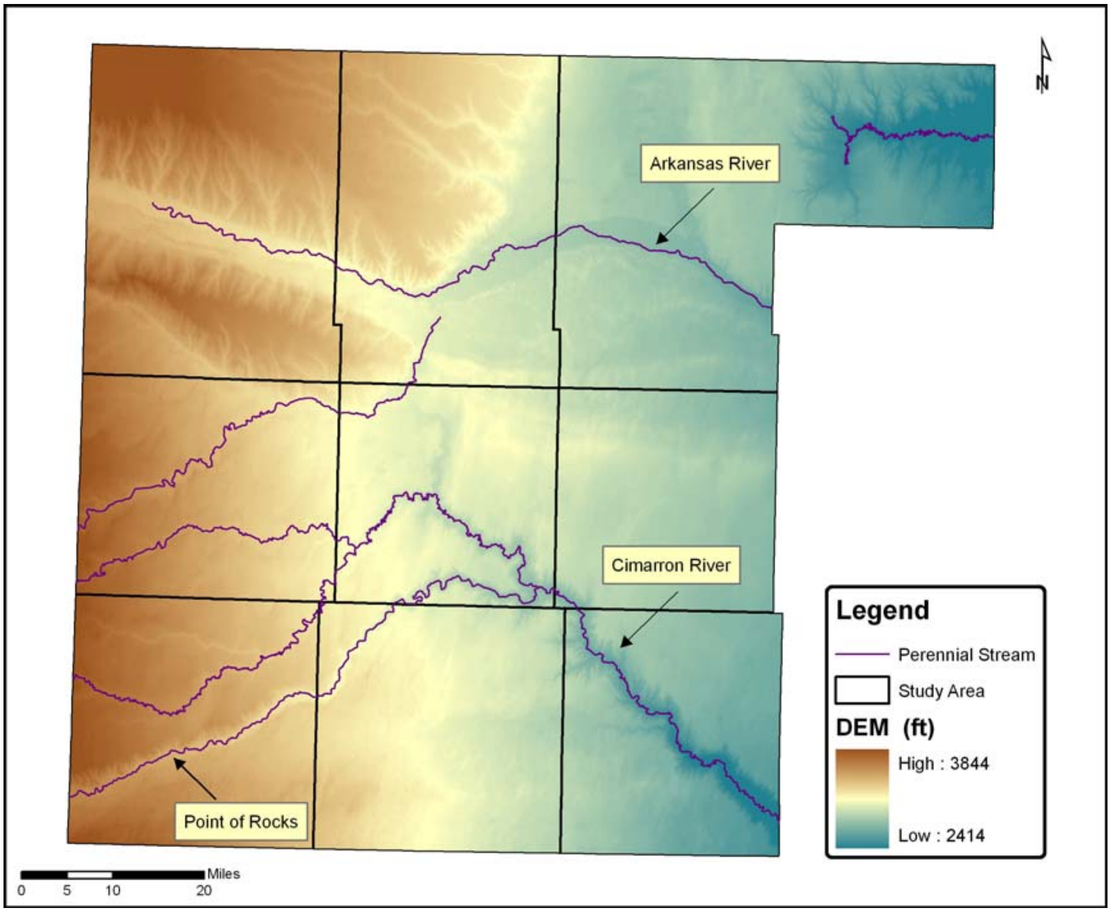


Figure 3.2: Digital Elevation Model (DEM) of the study area displaying the trends in surface elevation in the study area. The location of the two major river systems, Arkansas and Cimarron, are also displayed. Point of Rocks in Morton County is noted for reference.

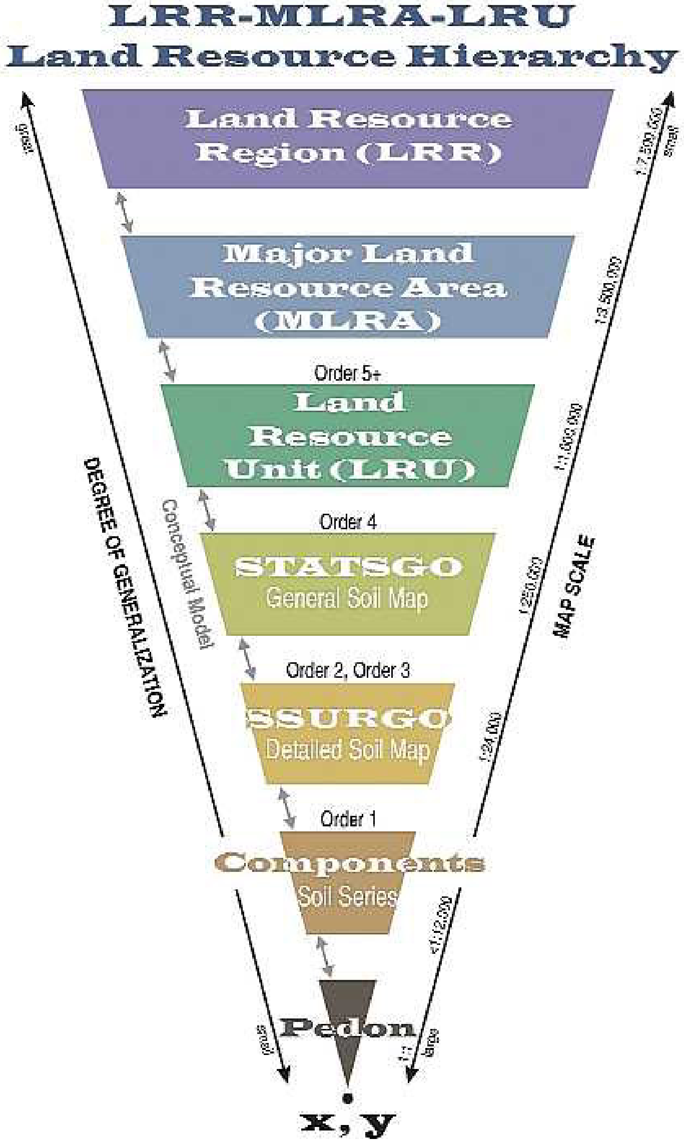


Figure 3.3: Graphical depiction of the MLRA classification scheme. The scheme utilizes a hierarchical approach in which a single soil unit, *i.e.* pedon, is classified into one of several different grouping based on the scale of analysis.

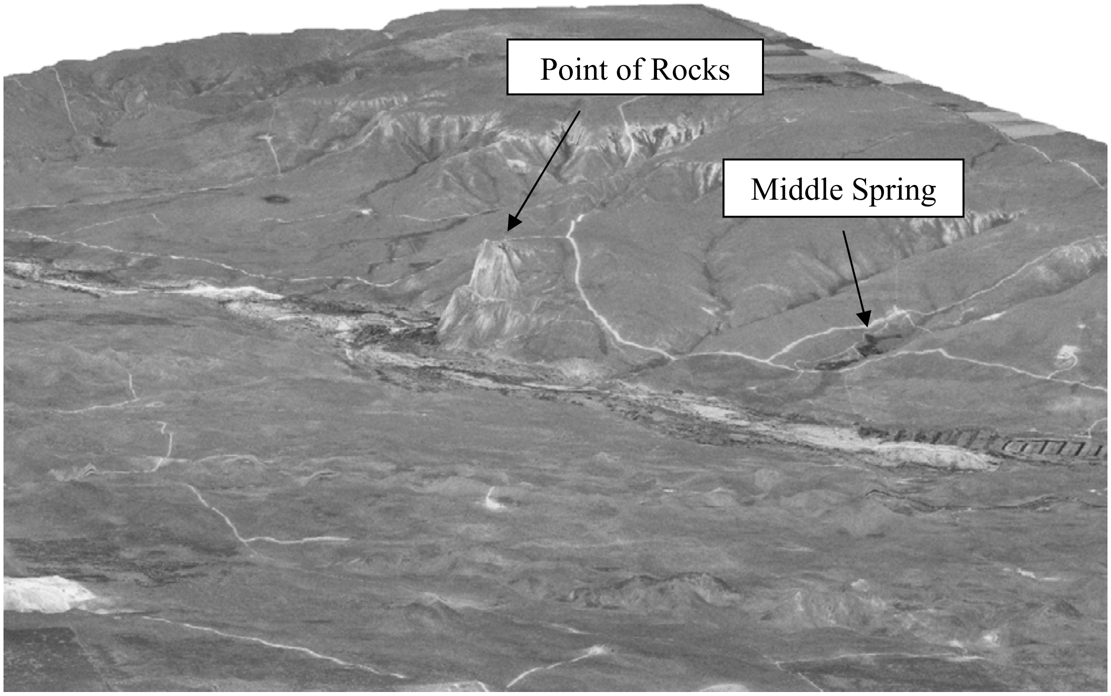


Figure 3.4: 3-dimensional oblique rendering of the Point of Rocks area of Morton County, Kansas. Note the close proximity of the Point of Rocks overlook and Middle Spring. Archaeological evidence indicates this area was heavily utilized by native populations prior to European settlement.

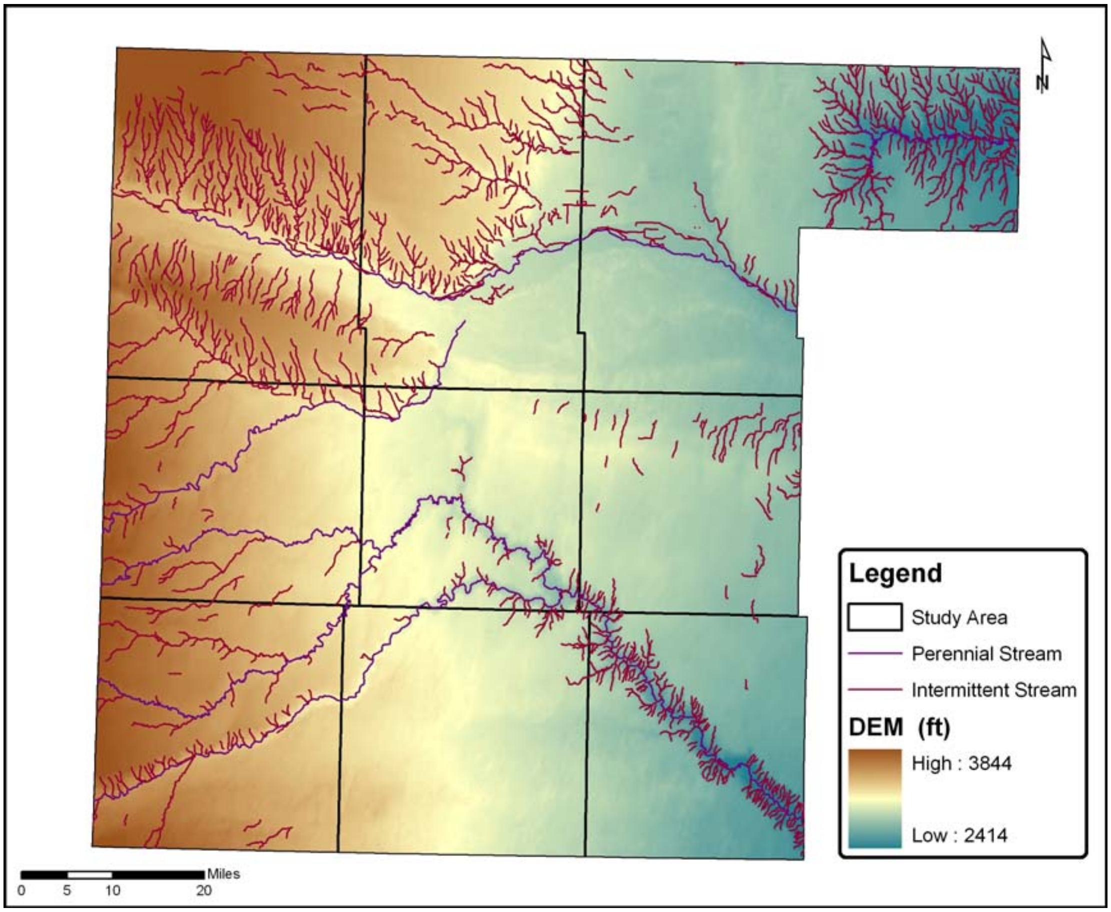


Figure 3.5: Hydrologic network within the study area. Stream segments are color coded based on Strahler stream order classification into ‘Intermittent’ and ‘Perennial’ flow regimes. Note the large expanse of upland areas with little or no reported surface water flow.

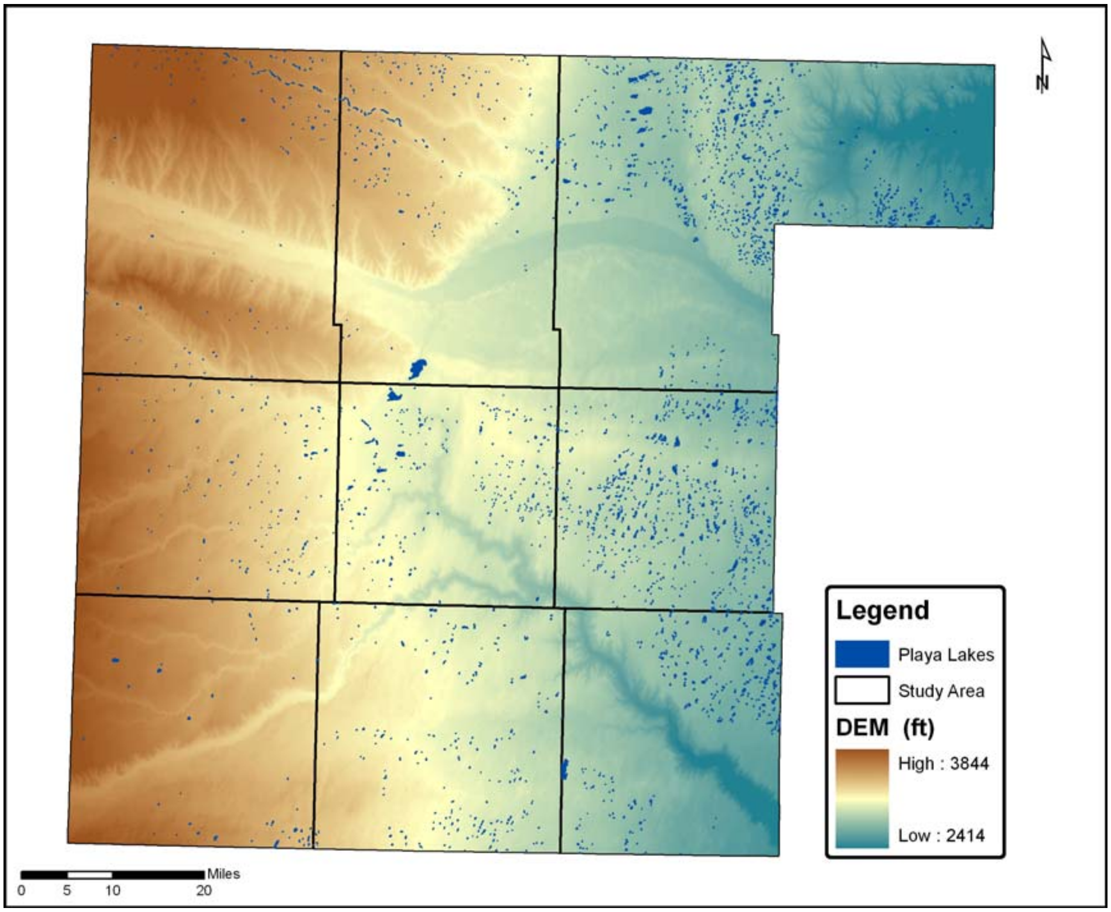


Figure 3.6: Location of playa lakes within the study area. Playa dataset constructed from SSURGO soils data. Note the density of playas in upland areas with little to no reported surface water flow.

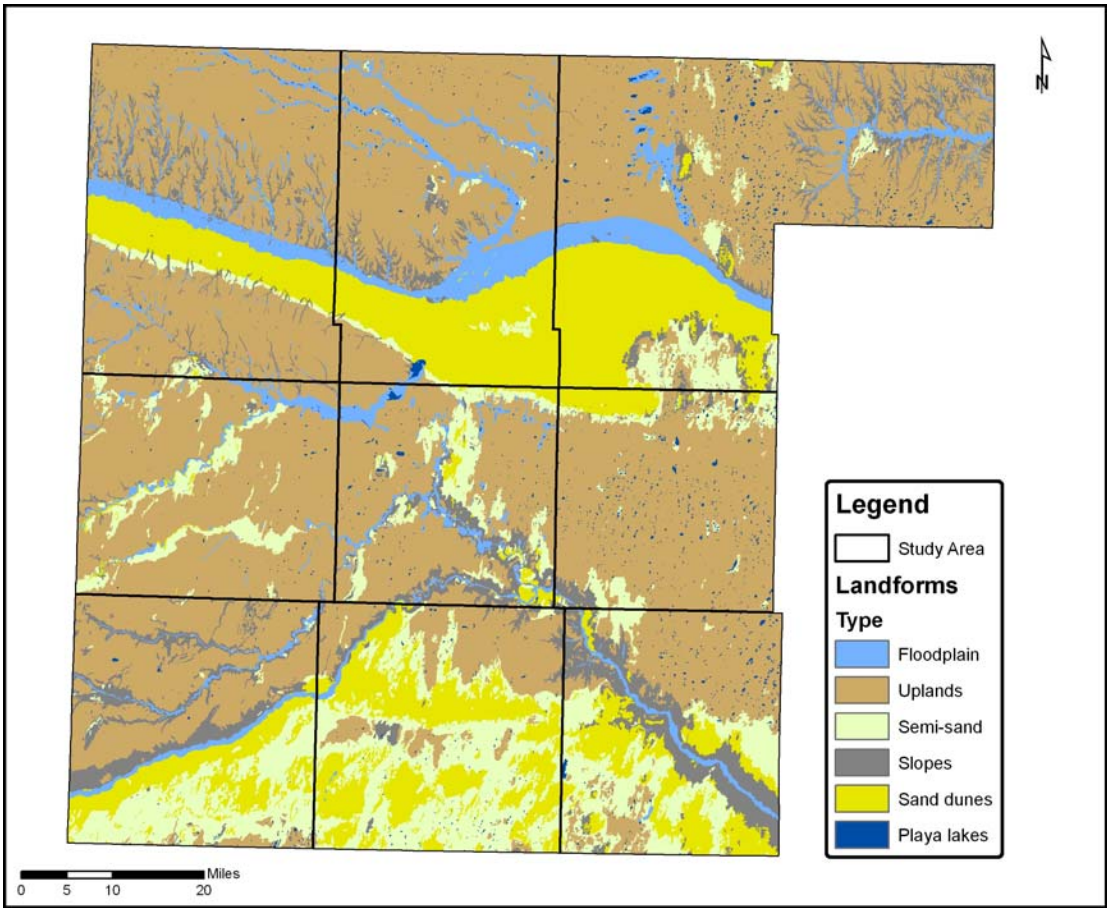


Figure 3.7: Landform map of the study area. Dataset created by generalizing SSURGO soils data; over 20,000 individual soil polygons were condensed into six potential landform categories.

Chapter 4

The Southwest Kansas Model

One goal of an archaeological predictive model is to classify a given land parcel into either an archaeology-present or archaeology-absent class based on measurable landscape characteristics, *e.g.,* distance to water, local relief, slope, etc. In this application, the statistical method used to determine the relationship between cultural material and the geographical characteristics of the location is binary logistic regression analysis. Landscape data for known site areas and known non-site areas were extracted from the GIS and entered into SPSS statistical software for analysis. Once the regression equation was developed, it was entered into the GIS and "mapped" across the landscape, meaning that all $30m^2$ grid cells contain an individual probability value computed from the logistic regression equation. Figure 4.1 displays the project workflow for the development of the Southwest Kansas APM.

Output of the logistic regression analysis was a continuous raster surface describing the probability of each land parcel for containing cultural material. Model development utilized a set of land parcels known to contain cultural material and a set of land parcels that do not contain cultural material. In order to assess model accuracy, the developed model was tested against a set of known archaeological locations that were withheld from the original model development. Ability of the model to predict the 'testing' sample of sites provides quantitative measures of the power or accuracy of the model.

Variables used in the model include *slope, distance to intermittent stream, distance to perennial stream, distance to playa lake, landform, relief within a 150-meter radius, relief within a 300-meter radius, relief within a 600-meter radius,* and a *'shelter index' within a radius of 150 meters*. Variable selection was designed to reflect components of the landscape significant for a hunter-gather subsistence strategy (Kvamme 1992). These variables are derivatives of modern map data that were created and stored within a GIS. Although landscapes change over long periods of time, this set of geographical variables was selected because they represent reasonably stable features during the last 15,000 years.

Spatial Database Construction

Generating the database used in model development required the compilation of archaeological and geographical spatial data from a variety of sources. Archaeological site location data was obtained from the Kansas State Historical Society (KSHS). The KSHS site registry is the official record of site locations for the State of Kansas and has been formatted for use in a GIS (Kansas State Historical Society 2002). Spatial data on elevation, hydrology, and soils were acquired to develop the environmental variables for the model. The Data Access and Support Center (DASC) maintains the web portal for GIS data distribution in the State of

Kansas and distributed all GIS data layers used for this project. Figure 4.2 displays a graphical representation of the database construction workflow.

As with all modeling applications, the quality of input data will affect the quality of the output. Incorrect input data, either environmental or archaeological, will adversely affect accuracy of the model output. Potential sources of error in archaeological data include the spatial position of site locations, the lack of reliable cultural affiliation for site materials, and poorly distributed sample data. Potential sources of error in environmental data include issues of map accuracy and the inappropriate use of geographic data sets.

Archaeological

Binary logistic regression analysis requires input data for both cases of the event class, site-present {S} and site-absent {S'}. The site-present event class consists of known archaeological locations, while the site-absent class should optimally consist of known non-site locations. For the site-present event class, the Kansas State Historical Society (KSHS) provided access to the archaeological site location database in February 2002; the KSHS GIS dataset consisted of a site location polygon layer and related attribute tables. The site location layer was formatted as an ESRI Arc/INFO polygon coverage for the entire state; attribute tables were dBase IV (.dbf) files organized on a county basis. Each polygon in the site coverage represents the field archaeologist's best attempt at gauging the activity area of a site. Since the majority of sites were recorded without the benefit of GPS, the precision of site boundaries is unknown. Polygon attribute tables of the site location coverage contained the official KSHS site number of each archaeological site, e.g. 14MT145. KSHS site numbers provided the database key for linking the site location coverage and attribute tables together in the GIS. The attribute tables contained detailed information about the location and archaeological content of each site. Data contained in the digital attribute tables were obtained from paper-versions of the official site submission forms and include information about site type, cultural affiliation, and descriptions of site setting and artifact content.

Several modifications were made to the archaeological site data to prepare it for the model. The first step was to clip the statewide polygon data to the boundaries of the study area (Figure 4.3). Second, attribute tables from each of the nine counties in the study area were joined to the polygon data. The final step involved filtering the site data to remove unsuitable sites from the input data. To limit the type of site information entering the model, site data were filtered to eliminate sites with a 'Historic' attribute in the cultural affiliation field or an 'Isolated Find' attribute in the site description field. Historic sites were removed so as not to dilute the model with spatial patterns unrelated to a pre-European existence, however sites classified as multi-component, indicating material from prehistoric and historic time periods, were retained. A site classified as an 'Isolated find' indicates the site consists of a single artifact, usually a single stone tool. A solitary artifact was not considered indicative of a settlement site and therefore not adequate for inclusion as an archaeological site in the model. These sites were removed so that only sites with multiple artifacts were used as input. Eliminating isolated finds refined the input data to only those sites at which significant activity occurred. The Mn/Model followed similar provisions for ‘Isolated Finds’ and ‘Lithic Scatters’ (Hudak et al. 2000).

After filtering, a total of 226 sites remained in the site-present event class. These were then randomly divided into two groups: the first group contained 2/3 of the total number of sites and was used to construct the model (n=151 sites), and the second was composed of the remaining 1/3 which were used to test the completed model (n=75 sites). Site data had to be converted from its original vector data structure to the raster data model. Site polygons were converted to the raster format and assigned a cell size (or land parcel size) of 30 m^2, equivalent to the finest resolution of the geographic data. In the raster format, the site-training group consisted of 7,917 cells (1,730 acres), while the site-testing group contained 3,344 cells (736 acres) (Figures 4.4 and 4.5)

As previously stated, the non-site samples are optimally based on a sample of land parcels in which cultural surveys have been performed and been verified not to contain archaeological material, in the study area however surveyed parcels were not equally distributed across all landforms. The goal of the binary logistic regression is to compare the environmental characteristics of known sites with the range of potential landscape choices. Therefore, the non-site sample needs to sample all possible landscape choices. Using this requirement as justification, the non-site sample was created using a random sampling method. A set of 3,900 points were randomly generated, buffered to 30 meters in diameter, and converted to the raster format with a cell size of 30m^2. In total, the non-site training sample contained 12,303 cells (2,706 acres) (Figure 4.6). The ratio of site to non-site training cells used in model construction is not agreed upon in the archaeological modeling literature; ranges between 1:1 and 1:10 have been reported with little statistical justification (Kvamme 1992; Warren and Asch 2000). For this project a ratio of 1:1.5 was chosen, with the final ratio of training pixels at 7,917:12,303 or 1:1.55.

Use of randomly generated non-site samples has been addressed by Kvamme (1992) and subsequently implemented widely in the archaeological modeling literature. A critical question concerning the non-site selection relates to the reliability of assuming that a non-site is actually a non-site. The non-site sample assumes that if a site is not reported at that location, then it is in fact a non-site. The inherent problem with this assumption is that it is impossible to know whether a non-site is actually a non-site without a survey. However, as Kvamme (1992) points out, archaeological sites are rare events and because of the low density of sites on the landscape, most likely a randomly generated non-site is in fact a non-site. A second set of randomly generated points was used to create the non-site testing sample. Once buffered and converted to the raster format, the testing sample contained 3,142 pixels. For the testing sample a 1:1 ratio of sites to non-sites was used, specifically 3,344:3,142.

Limitations of Archaeological Database

The majority of sites reported in the site database contained lithic tools, mostly non-diagnostic lithic scatters. Lithic materials are inherently difficult to use for determining cultural or temporal affiliation; only artifacts with diagnostic physical traits can be effectively dated or assigned to a specific cultural group. Site function is also difficult to discern on the basis of lithic scatters. Because of this limitation the information contained in the cultural affiliation database field has limited utility for use in an APM. The lack of reliable cultural affiliation and site function information dramatically impacts the type of predictive modeling the data could be used for; because the attribute data lacks specificity, the binary classification of site-present and site-absent was chosen for the model. Similar procedures were followed in the Mn/Model (Hudak et al. 2000).

One drawback to the structure of the attribute database involved the descriptive data fields. These fields contain descriptions of site setting and artifact content derived from the KSHS site forms. Descriptions from the site forms were directly copied into a series of consecutive data fields. For example, each text-based data field might contain space for 50 characters; if the site description was originally one paragraph and contained 300 characters, the description would span six consecutive data fields in the attribute database. Information in these fields was not structured to utilize standard database operations, including sort and query, and was subsequently of little utility to the model. Potential improvements to the KSHS database would be a standardization of these data fields.

The majority of sites in the database were either surface finds or shallowly buried. Temporally most sites are classified as *prehistoric* with some *ceramic* and *paleoindian*. Specifically, 204 of the 226 sites used in the model are classified as *prehistoric*, 20 as *ceramic*, and 4 as *paleoindian*. Field survey projects that generated the majority of data found in Morton County were conducted by the University of Kansas in 1975; Brown (1977) indicates that based on the presence of arrow points at the majority of sites in Morton County, these sites postdate A.D. 1. In the KSHS database, these sites were given a 'prehistoric' designation. Assigning sites to a specific cultural affiliation, *e.g.* Keith Variant or Dismal River Aspect, was not possible due to the lack of diagnostic features. Therefore, because the majority of input data were 'prehistoric', the model output is biased towards surface finds deposited during the last 2,000 years. While this might be considered problematic for predicting the locations of sites older than the Woodland period, many hunter-gather sites demonstrate repeated habitations thereby indicating the general utility of some landscape locations. The identification of landscape 'hot-spots' of desirable environmental characteristics should increase the detection of materials and sites.

Environmental

Environmental variables are used as the independent or predictor variables for the probability model. Variables were selected on the basis of available digital data and potential for influencing hunter-gather site location. The goal of variable selection is to identify the smallest set of geographic factors that could influence site selection. Additionally, it is important that data sources represent the ‘universe’ of the variable and have numerous occurrences within the study area. For example, data about naturally occurring springs might be useful in the model, however, the available spring data may not account for all spring locations. Also, if the total number of springs were low, then including the springs data as a variable would diminish the predictive power of other variables or caused spurious correlations.

Additional variables that could have been useful but were not used in the model include data on floral and faunal resources. Climatic fluctuations in the last 15,000 years have inevitably affected the distribution of floral resources, and subsequently affected faunal communities dependent upon the vegetation. Due to the difficulty of modeling these variables through time, no specific floral or faunal variables were used in the model.

Due to the constraints on variable inclusion, the final variables used in the model reflect a set of geomorphic variables that demonstrate a relative stability through the extent of human occupation of the area. Variables were grouped on the basis of elevation derivatives, water resources, and landform data. Table 1 lists the variable name, source of data, data developer, and whether the data came directly from a primary source or was derived from the primary source.

Table 1: Environmental Variables

	VARIABLE	SOURCE	AGENCY	TYPE
Terrain	Slope	DEM	USGS	Derived
	Relief (150m, 300m, 600m)	DEM	USGS	Derived
	Shelter Index (300m)	DEM	USGS	Derived
Water Resource (distance from…grids)	Intermittent Stream	SWIMS dataset	KDHE (Kansas Department of Health and Environment)	Derived
	Perennial Stream	SWIMS dataset	KDHE	Derived
	Playa Lakes (w.90m buffer)	SSURGO	NRCS	Derived
Geomorphology	Landform	SSURGO	NRCS	Derived

Elevation

The USGS Digital Elevation Model (DEM) formed the basic unit of analysis, *i.e.* land parcel size, for the project; due to the number of derivatives from the DEM, the $30m^2$ pixel size is used as the standard raster grid cell size for the model. Raw elevation values are of limited utility in the study area due to the large size of the area and the general slope of the land from west to east. However, elevation values can be transformed into meaningful information for describing the landscape condition in a local area. The DEM was used to calculate several derivatives including slope, relief, and a 'shelter index'. Each of these variables can be shown to have an impact on human activity patterns and represent a logical choice for inclusion in the model (Butzer 1982; Jochim 1976; Kvamme 1992).

Slope was created using the standard routine within ESRI's Spatial Analyst extension for ArcGIS 8.3 (Figure 4.7). The multiple relief measures and the shelter index were generated using focal (neighborhood) functions and implemented using the Raster Calculator within Spatial Analyst. Relief was calculated by determining the range of elevation values (range = maximum – minimum value) within a given neighborhood. It was assumed that relief played an important role in site location, however the size of the neighborhood in which to calculate the relief value was unknown, therefore three relief measures were calculated with radii of 150m, 300m, and 600m respectively (Figure 4.8).

‘Shelter index’ is a metric designed to be a measure of how ‘sheltered’ or ‘exposed’ a land parcel is with respect to its surrounding environment. The idea behind the ‘shelter index’ is to calculate the internal volume of a cylinder of known size that is placed over a land parcel and its neighborhood (Kvamme and Kohler 1988). Volume of the landscape surrounding the land parcel is calculated and subtracted from the volume of the cylinder. If the volume is relatively large, the land parcel is located on a hilltop and is exposed to the surrounding landscape. If the volume is relatively small, the land parcel is located in a valley bottom and is sheltered from the surrounding landscape. The shelter index was calculated with a 300m radius, which was assumed would provide a large enough area for considering the local affects of topography (Figure 4.9). Specifically, the Arc/INFO code used to develop the shelter index is listed below (Kvamme and Kohler 1988):

```
// create a layer (equal to the input DEM) with all values equal to 1
allCells = ([dem] * 0) + 1

//calculate the area of a circle with a radius of 300 meters
temp1 = focalsum ([allCells = 1], circle, 10, data)

//multiply the area of the circle by the DEM +20 meters to compute the cylinder volume
temp2 = ([dem]+20) * [temp1]

//calculate the volume of the DEM with a 300 meter radius
temp3 = focalsum ([dem], circle, 10, data)

//subtract the volume of the DEM from the volume of the cylinder
shelter300m = [temp2] - [temp3]
```

Hydrology

Water is a critical factor in human settlement of the High Plains (Wedel 1963; Holliday 1997) and in hunter-gather peoples (Butzer 1982; Jochim 1976). Accordingly, the Surface Water Information Management System (SWIMS) dataset, generated by the Kansas Department of Health and Environment, provided the hydrographic source data for the project. Due to the aridity of the study area during the last 8000 years, water resources needed to be stratified on the basis of availability. Two major river valleys are located in the study area (Arkansas and Cimarron) along with numerous smaller streams (Figure 3.5). In order to recognize the potential difference in flow regimes between the large perennial rivers and the smaller intermittent streams, the SWIMS dataset was separated into two groups on the basis of Strahler stream order. All stream segments with a Strahler classification of 4 or greater were considered perennial water sources, while those classified as 3 or lower were considered intermittent streams. This determination was based on published literature in other areas of the Great Plains (Warren and Asch 2000) and personal knowledge of the study area.

The location of stream segments is of limited utility for predictive modeling. In its native format, vector representations of stream locations only provide a binary condition, *e.g.,* water or no water. In order to be useful, the data were converted to a continuous data surface describing the distance of each individual land parcel to the closest water source. The Spatial Analyst extension of ArcGIS was used to generate 'distance from….' grids for both the intermittent and perennial streams layers (Figure 4.10 and 4.11).

Another source of water on the High Plains comes from playa lakes, or intermittent ponds that occasionally fill with water during large rain events (Holliday 1997). Using soils data as a base, Johnson and Campbell (2004) created a GIS database of playa locations by extracting the diagnostic soils that occur in the basins of playa lakes. Once the vector polygon boundaries of playa soil bodies were extracted, a 90-meter buffer was placed around each basin. Previous research has shown that playa extent is difficult to determine on the basis of soils alone and the buffer is intended to represent the activity area associated with a playa. Once buffered, a 'distance from…' routine was used to create a 'distance from playa' data layer (Figure 4.12).

Landforms

A visual analysis of the spatial distribution of sites clearly indicated a preference for certain landforms. SSURGO data created by the Natural Resources Conservation Service (NRCS) provides the best continuous digital data available about the composition of the land surface. SSURGO data represent the finest resolution of soils data available; mapped at 1:24,000 scale, the building blocks of SSURGO data are soil units. According to the Major Land Resource Area (MLRA) soil mapping standards (Soil Survey Staff, 1981), an individual soil unit occurs only in a specific landscape/geomorphic position. Because specific soil units occur only in designated landscape positions, it is possible to generalize the specific soil units into a geomorphic map. The reliability of SSURGO for this application is better in Kansas, and other agricultural states because of the economic imperative for quality soils information. Kansas is one of the few states in the country that has SSURGO level data for every county in the state.

The goal of the landform data was to generalize the more than 200 different soil units found within the study area into a geomorphic map consisting of only six classes, specifically floodplain, steep slopes, upland, semi-sand or sand sheets, sand or sand dunes, and playa lakes. Reclassifying the soil units was a manual process that required interpreting the landscape position of each soil unit from the NRCS documentation. In total, 20,000 individual soil polygons were classified into the landform categories. The six final classes represent a generalized and realistic set of landforms of the area (Figure 4.13). As with the ‘distance from playa’ variable, playa lake soils units were buffered to 90 meters and then converted to the raster format. In order to visualize the detail of the SSURGO data, additional graphics were created for Morton County. Figure 4.14 displays landform data in its generalized state, and Figure 4.15 displays landforms with the SSURGO polygons.

Data Extraction

Once the GIS database was built, environmental data for site and non-site locations were extracted and exported to SPSS (Statistical Package for Social Scientists) software for analysis. Using the site and non-site training pixels as a sampling mask, data were extracted for each of the independent variables using the SAMPLE command in the Raster Calculator (Figure 4.16). Output of the SAMPLE command is a tab-delimitated text file that was imported into SPSS and compiled into a single dataset. For each land parcel, the archaeological condition and associated values from the environmental variables are written out to an individual row; therefore 7,917 site-present and 12,301 site-absent rows of data were extracted. Once compiled in SPSS, the data were ready for statistical analysis.

Model Construction

In order to determine if the proposed environmental variables should be included in the model, univariate statistical comparisons (Mann-Whitney U and Komolgorov-Smirnoff) were used to determine if the two archaeological event classes had significant differences between them for each continuous variable. Due to the large number of samples, the tests were performed three times, first with all the samples, second using a 10% sample, and finally using a 1% sample. The sub-samples were created within SPSS and are random samples. In the first run, using all samples, all variables were found to be significantly different at α=0.005. However, the effect of the large sample size should artificially lower the significance values. The 10% sample used in the second run displayed the same pattern, all variables significant at α=0.005. For the 1% sample, all variables were significantly different at α=0.05. Significant differences for all the variables indicate that each of the environmental variables is appropriate for inclusion in the model, necessary for the overall validity of the model. An effort was made to use the smallest set of explanatory variables possible.

While the landform variable was not statistically tested, a comparison of the distribution of landforms for the entire study area and known sites clearly indicate a preference for particular landforms. Specifically the ‘Slopes’ category is differentially selected for (7.5% total landscape versus 39% of site locations) and the ‘Uplands’ are selected against (56% total landscape versus 15% of site locations). Based on the large variation in landform percentages between the overall landscape and the location of archaeological sites, it was determined that landforms were a significant variable in site selection and the use of a landform variable in the model is justified. Table 2 contains numerical data on the distribution of landform data in the different event classes.

Table 2: Landform Distribution

Landform Distribution		All Sites		Non-Site (training)	
6.0%	Floodplain	9.1%	Floodplain	5.9%	Floodplain
55.8%	Upland	14.7%	Upland	56.2%	Upland
13.0%	Semi-Sand	9.6%	Semi-Sand	12.5%	Semi-Sand
7.5%	Slopes	39.1%	Slopes	6.7%	Slopes
15.5%	Sand	27.0%	Sand	16.7%	Sand
2.2%	Playa	0.5%	Playa	2.0%	Playa
20,440,315	Land parcels	11,261	Land parcels	12,301	Land parcels

In terms of statistical analysis, the specific method chosen for model construction was a backward, step-wise binary logistic regression. Initially all independent variables are used in the equation and the power of the model is calculated; next, each independent variable is iteratively removed and the power is recalculated. If the change in model power is significant, the variable with the least explanatory power is removed from the set of independent variables and the process of power calculation and variable removal is repeated. Processing continues until the removal of a variable does not significantly change the power of the model. Once completed, the remaining variables all have significant explanatory power (Clark and Hosking 1986).

This specific model was also run in a forward stepwise method, however, the difference with the backward step-wise method was negligible. In the backward stepwise model only the 'Relief within 300m' variable was excluded from the analysis. In the forward stepwise method both 'Relief within 150m' and 'Relief within 300m' were excluded. Considering the insignificant difference between these two approaches it was determined the backward stepwise model was a more logical approach because all the environmental data layers are used initially as explanatory variables. From an archaeological perspective, it reasonable to assume that site selection was based on a simultaneous evaluation of multiple environmental criteria; this is best represented statistically in the backward stepwise method. Additional discussion of the internal statistical metrics is not required for two reasons. First, the statistical metrics used to determine the effectiveness of the model construction are within reasonable limits, and second, assessment of model performance is conducted using an independent testing sample. Specific details of the regression model construction are included in Appendix A in the form of SPSS output tables.

It is important to note that the model described herein (termed Model 8) is the eighth of ten iterations run on the dataset. Model 8 represents one of the simplest to understand and is, statistically speaking, the most powerful of the ten models. Variations to the modeling approach included whether to use only site centroids as data input, different random configurations of site samples, variations in the site-training / site-testing ratio, and using Morton County sites as the site-training and all other sites as site-testing. Ultimately Model 8 made the most theoretical sense and statistically performed the best.

Model Output

The regression equation developed within SPSS is mathematically written as follows:

$$Z = -2.701459 + (\text{Dist. to Inter.} * -0.000328) + (\text{Dist. to Perr.} * -0.000053) + (\text{Dist. to Playa} * 0.000196) + (\text{Floodplain} * 0.005184) + (\text{Upland} * -0.969463) + (\text{Semi-Sand} * 0.509768) + (\text{Slopes} * 1.535854) + (\text{Sand} * 0.867705) + (\text{Relief150} * -0.009843) + (\text{Relief600} * -0.012883) + (\text{Shelter Index} * 0.001617) + (\text{Slope} * -0.025225)$$

In order to convert the regression output to a probability score the following equation is also required:

$$\text{Prob(S)} = 1/(1+\text{EXP}(-Z))$$

This equation represents the best quantitative description between the occurrence of archaeological sites and the environment developed for the study area. For any given location, the environmental conditions inherent to that location can be entered into these two equations and the output is a numerical value, constrained between 0 and 1, which describes the potential of that location to contain archaeological material. A location with a score near 0 indicates a set of environmental characteristics unlike those found in the site-present class (or characteristics similar to the site-absent class), while a score near 1 represents a location with characteristics similar to those in the site-present class.

Once calculated, the regression equation is applied to every 30 m^2 land parcel in the study area. GIS methods, referred to as 'Map Algebra' or 'Cartographic Modeling', are used to implement the equations (Tomlin 1990). The equations were re-entered to the GIS using the Raster Calculator tool within Spatial Analyst. The GIS calculates the output of the regression equation for every land parcel or raster cell in the study area, which in this case is over 20 million parcels. The resulting output is a decision surface of continuous data values containing the probability score for each land parcel in the study area. See Figure 4.17 for an image of the final model.

In terms of visual analysis, the spatial pattern generated by mapping the output equation reveals some interesting landscape patterns. From a macro perspective, the dominant feature of the probability surface is the elevated values along the major hydrologic drainages and the low values within the large upland areas. However, viewing the entire study area masks the detail of the model. There are over 20 million land parcels that are evaluated on an individual basis, resulting in a rich spatial pattern that only becomes apparent when viewed at a finer scale. Figures 4.18 and 4.19 display model results focused on Morton County at progressively finer resolutions. When zoomed into a scale of 1:50,000 or larger, the unique computation of each land parcel becomes apparent. Examples include the variability of adjacent pixels within small drainages and the sand dunes and sand sheet areas.

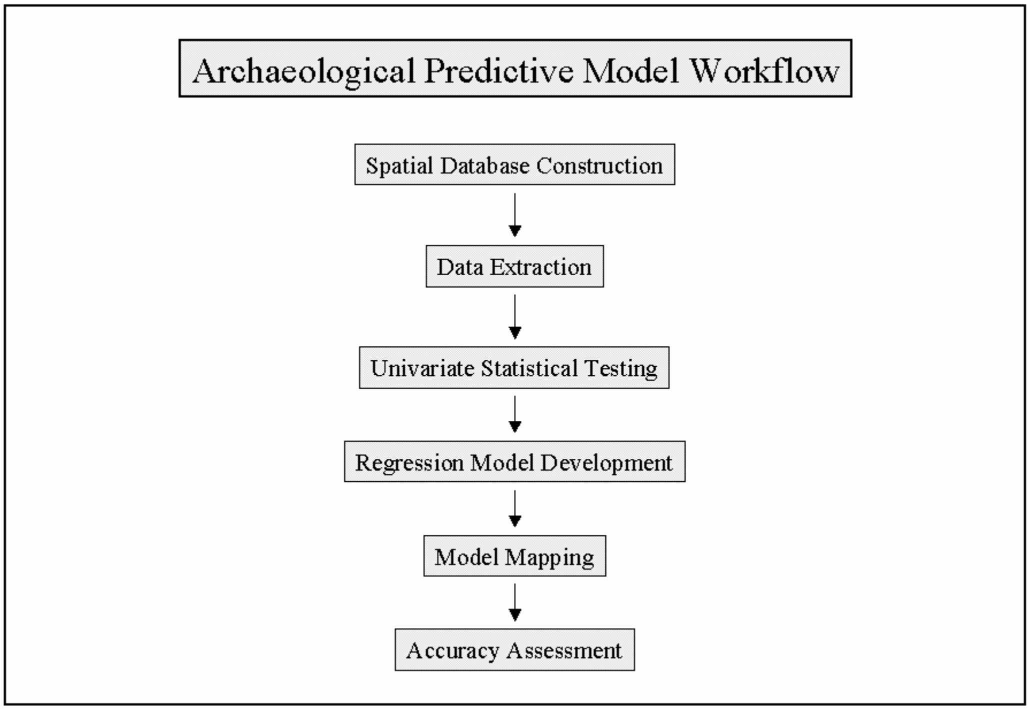

Archaeological Predictive Model Workflow
Spatial Database Construction
Data Extraction
Univariate Statistical Testing
Regression Model Development
Model Mapping
Accuracy Assessment

Figure 4.1: Processing steps for the development of the Southwest Kansas APM.

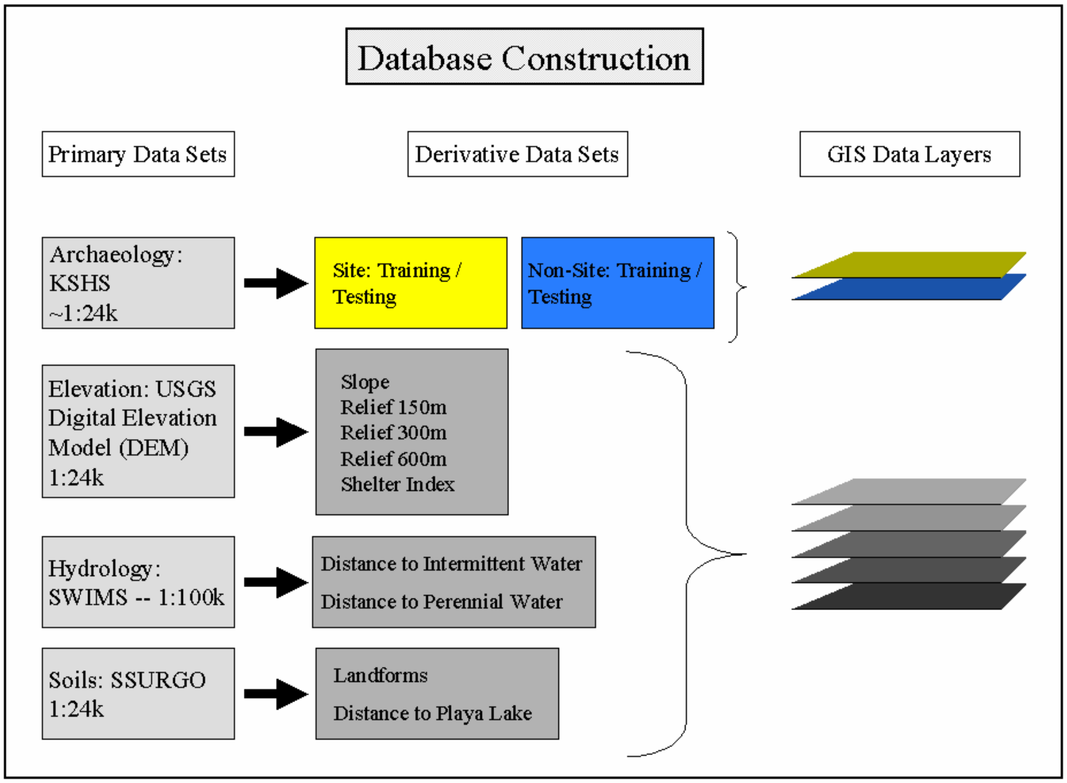


Figure 4.2: Workflow for the creation of the GIS database used in the APM. Note that each of the final datasets are derivative products of some primary data source.

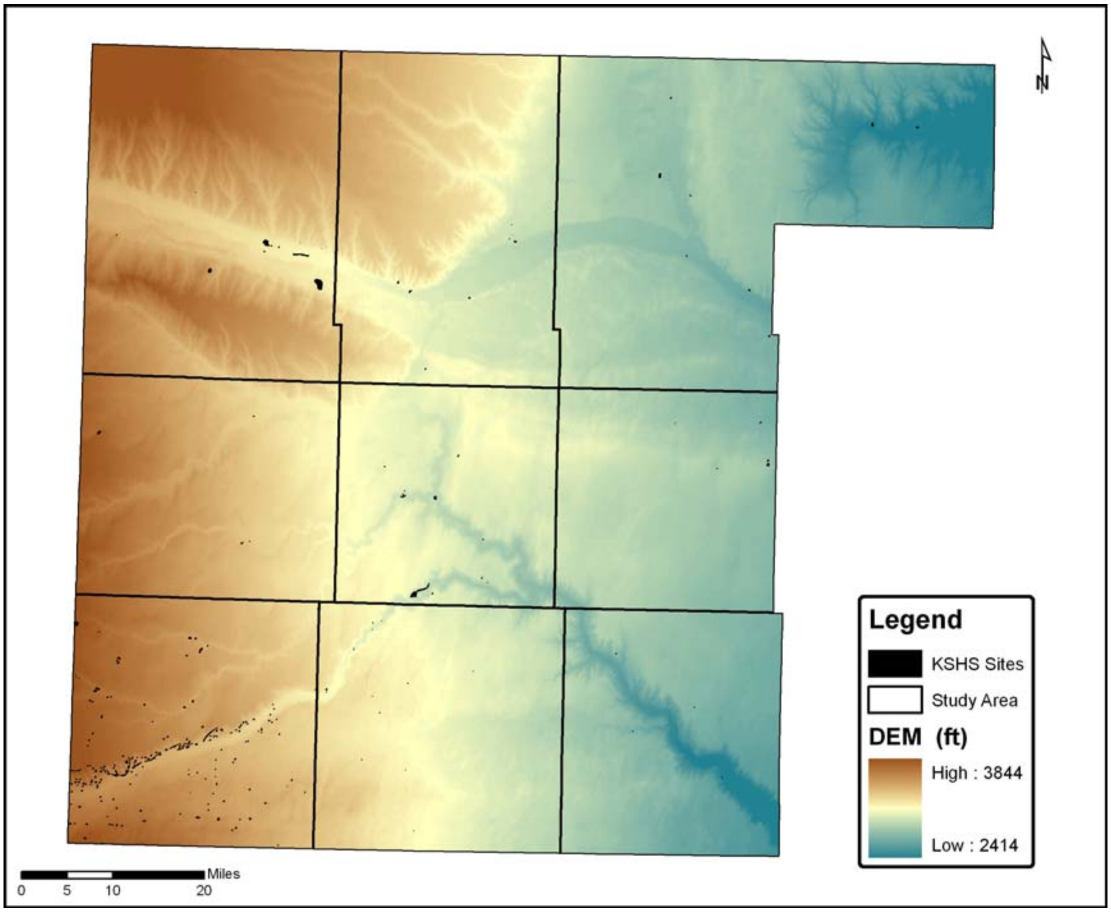


Figure 4.3: All archaeological sites in the KSHS database.

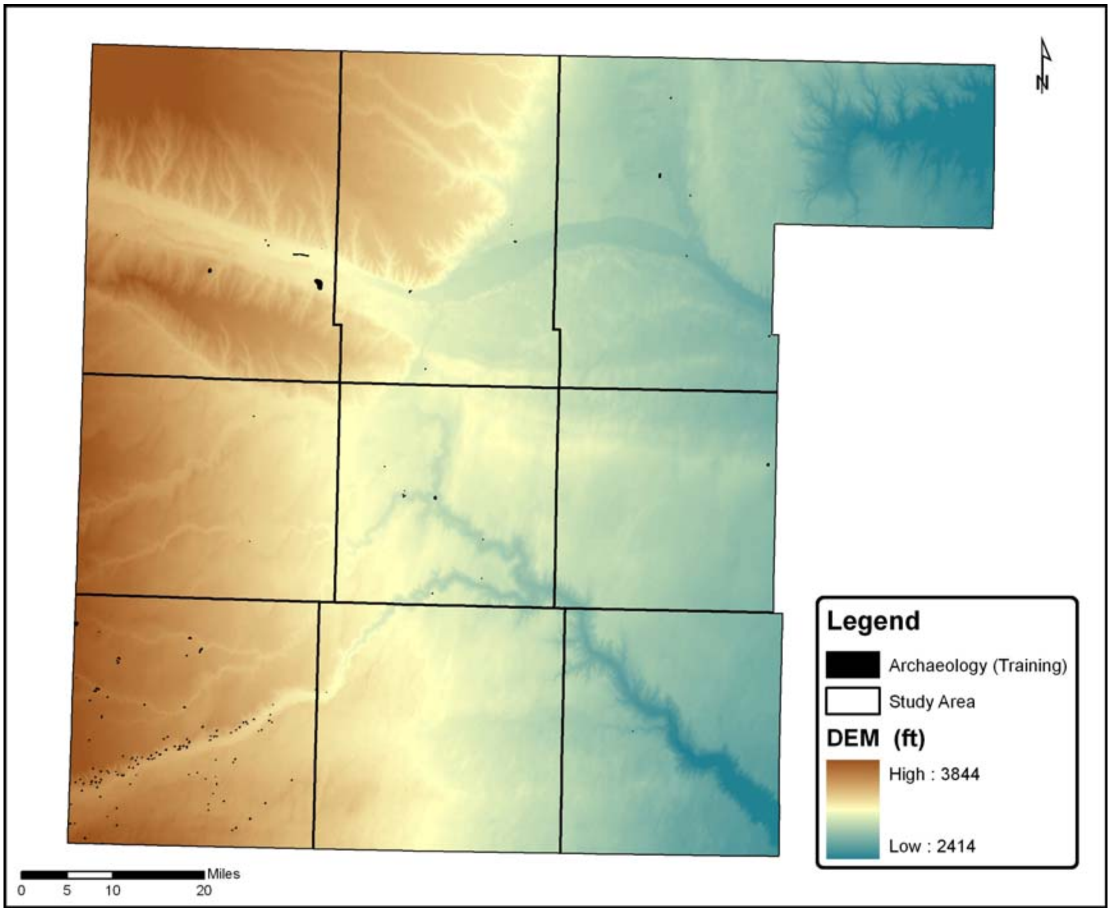


Figure 4.4: Archaeological sites used to train the archaeological predictive model.

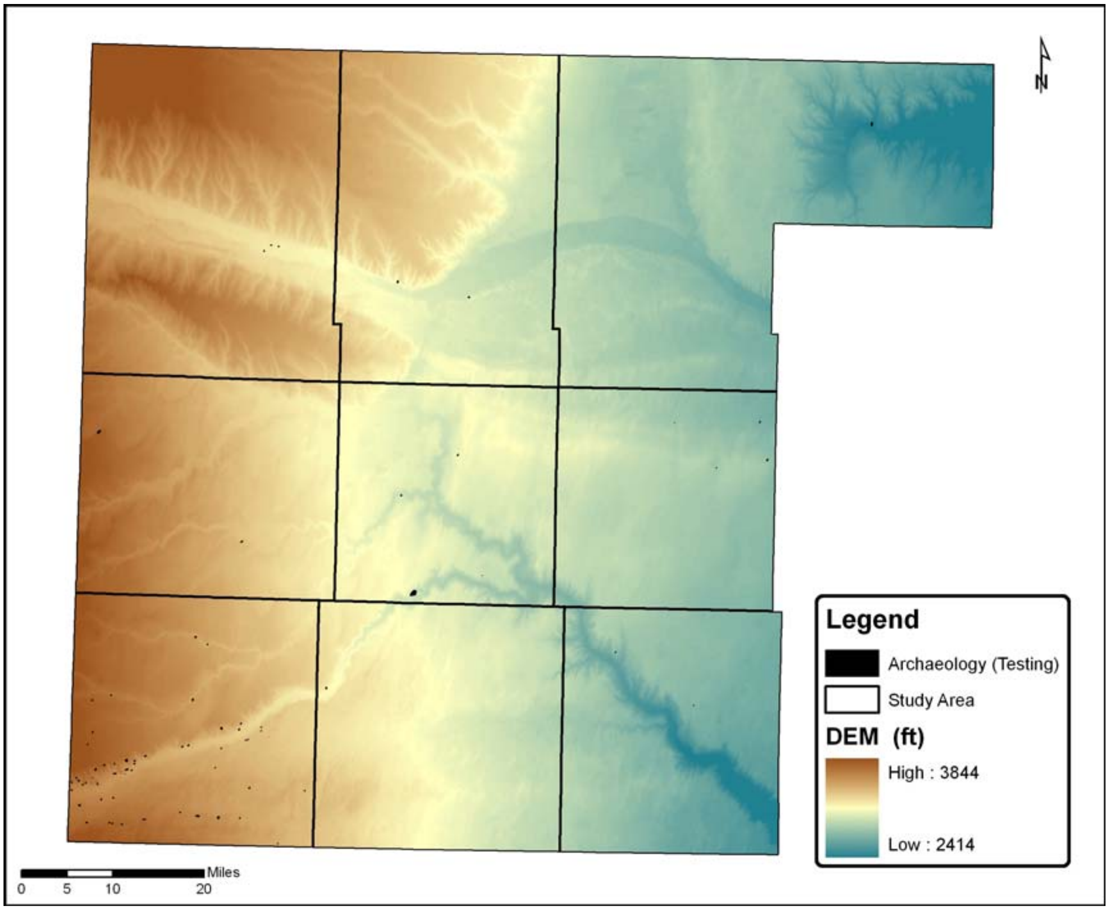


Figure 4.5: Archaeological sites used to test the accuracy of the archaeological predictive model.

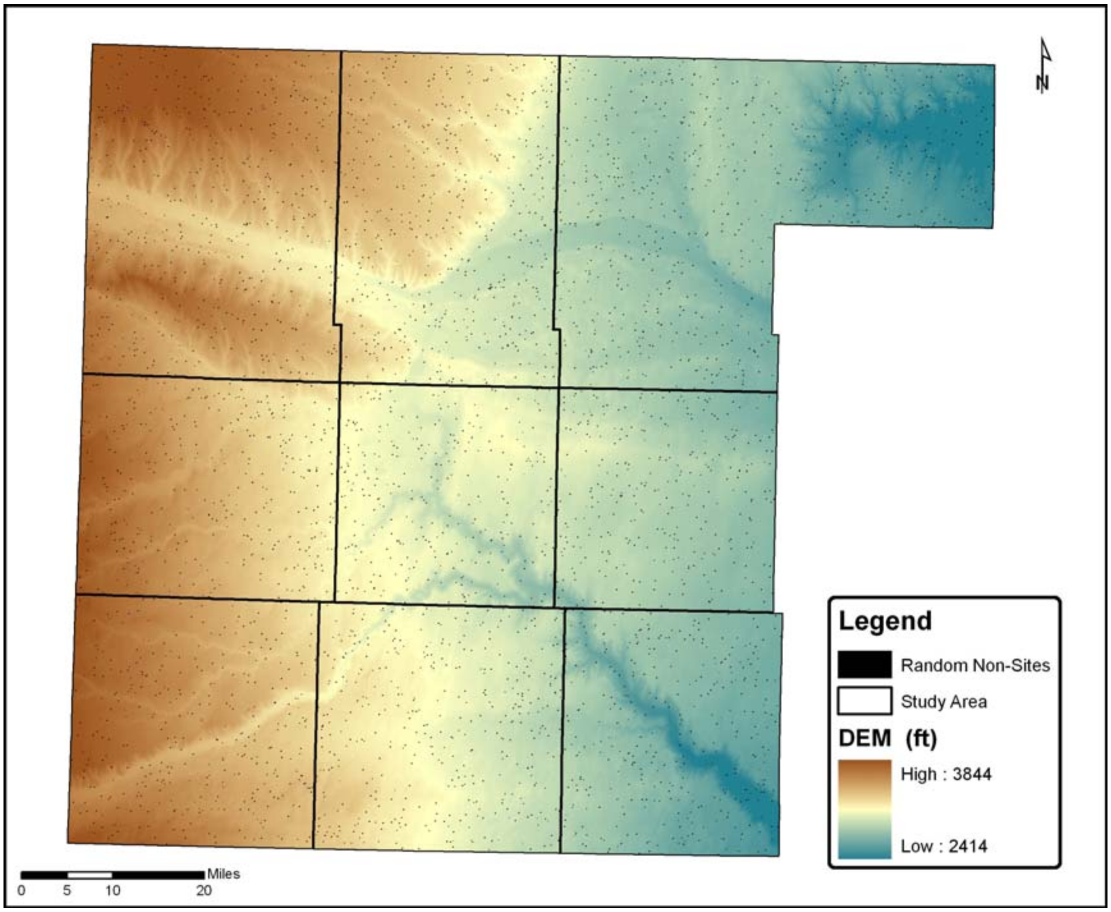


Figure 4.6: Randomly generated ‘Non-sites’ used in model construction and model testing.

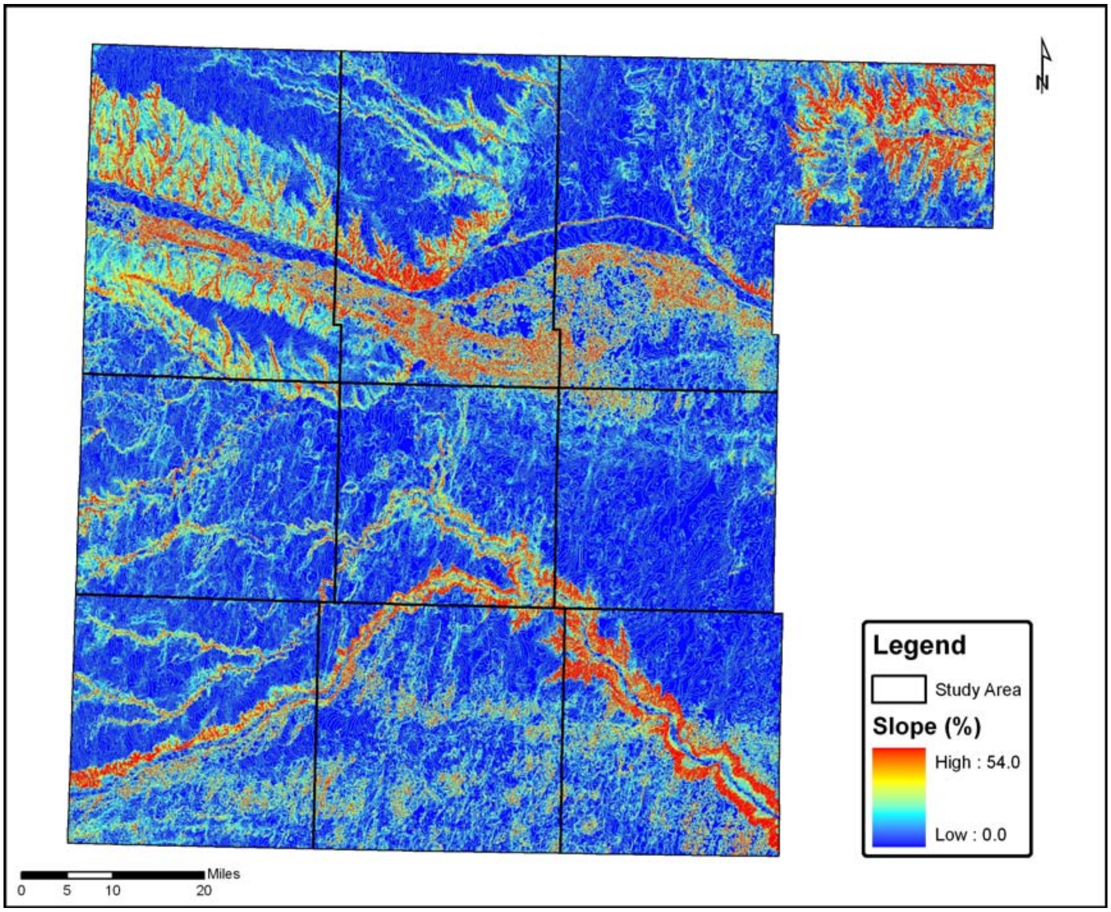


Figure 4.7: ‘Slope’ variable generated from the digital elevation model (DEM).

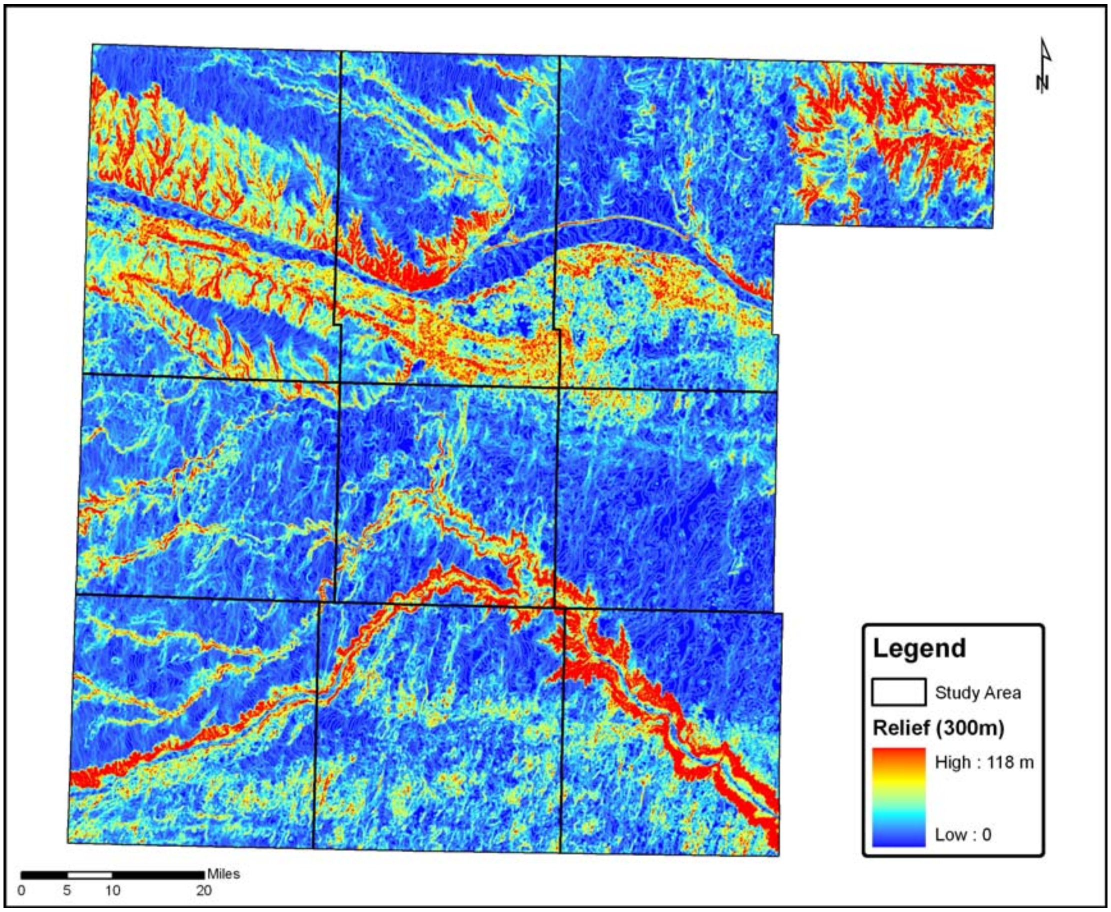


Figure 4.8: ‘Relief with a 300-meter radius’ variable generated from the DEM.

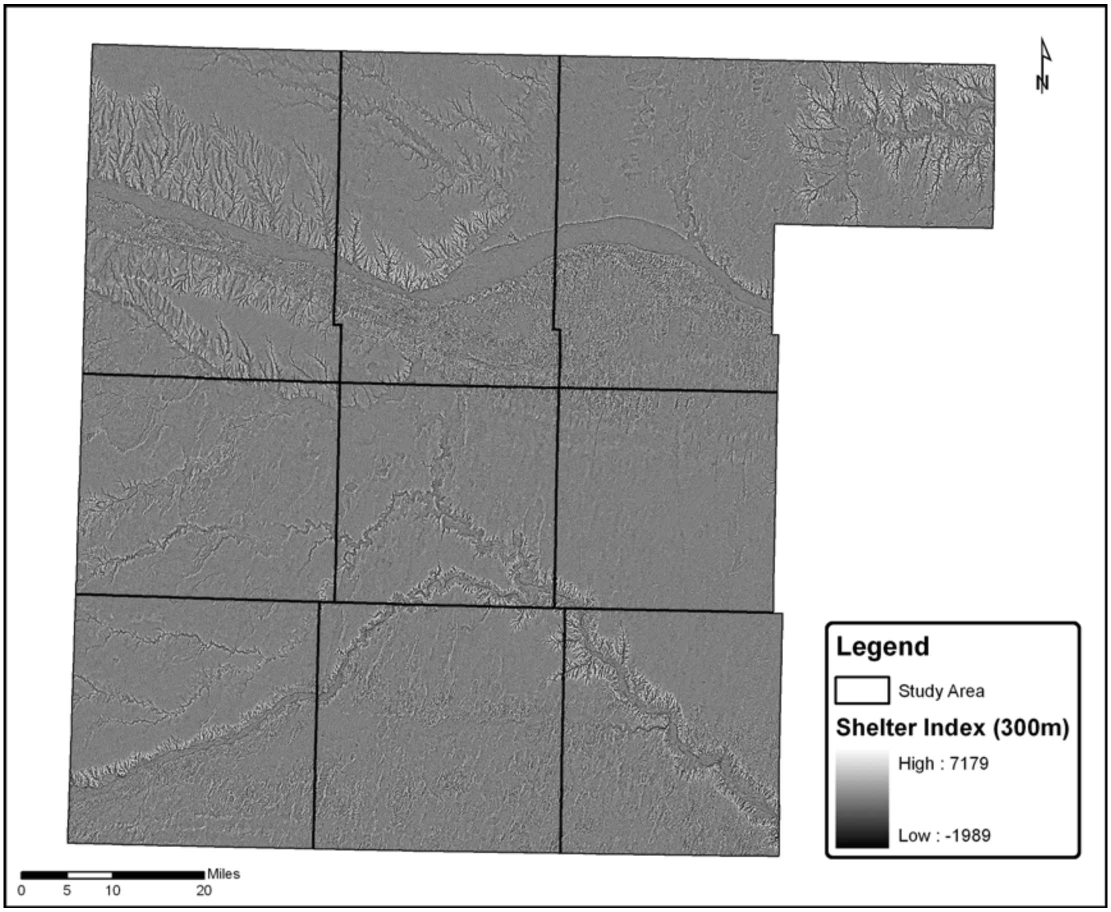


Figure 4.9: 'Shelter Index' variable (300m radius) generated from the DEM. High values indicate a land parcel is exposed, low values indicate a land parcel is sheltered.

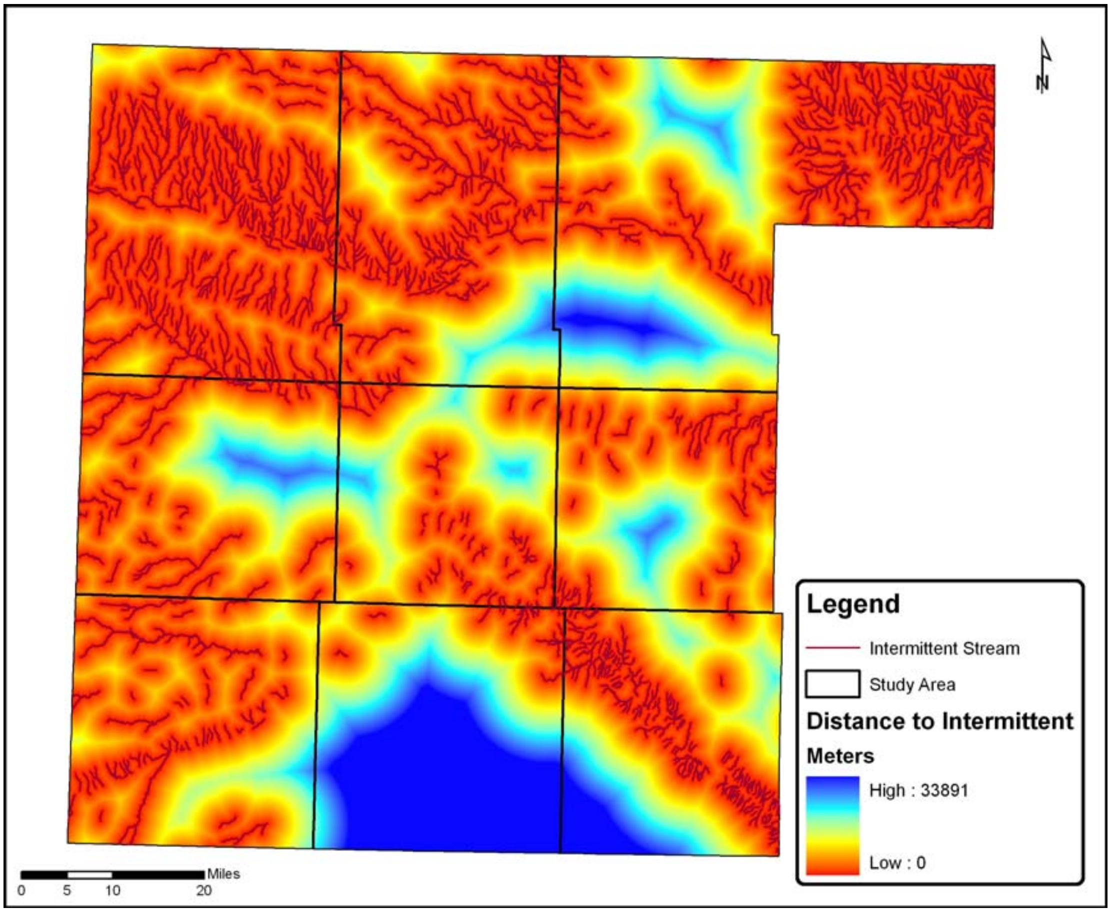


Figure 4.10: 'Distance to Intermittent Stream' variable generated from the SWIMS dataset. The value of each land parcel represents the distance to the closest intermittent stream segment.

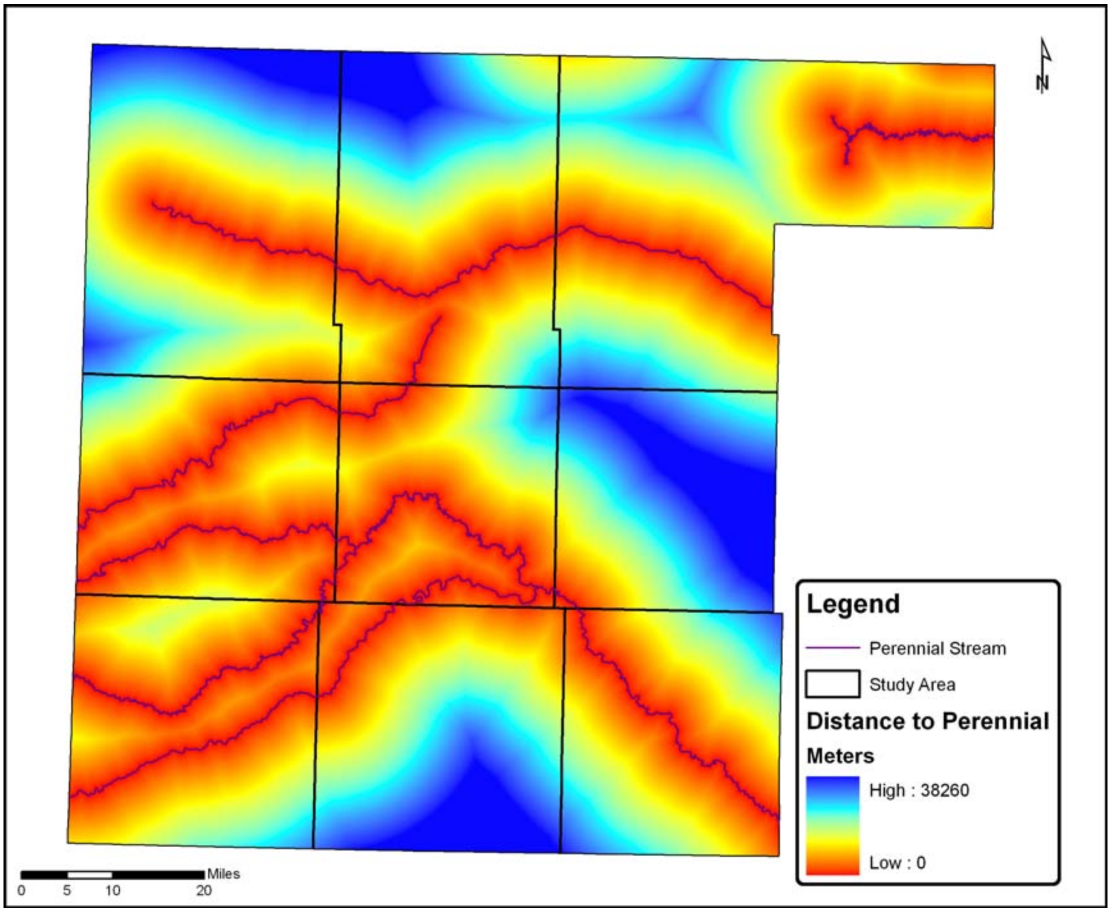


Figure 4.11: 'Distance to Perennial Stream' variable generated from the SWIMS dataset. The value of each land parcel represents the distance to the closest perennial stream segment.

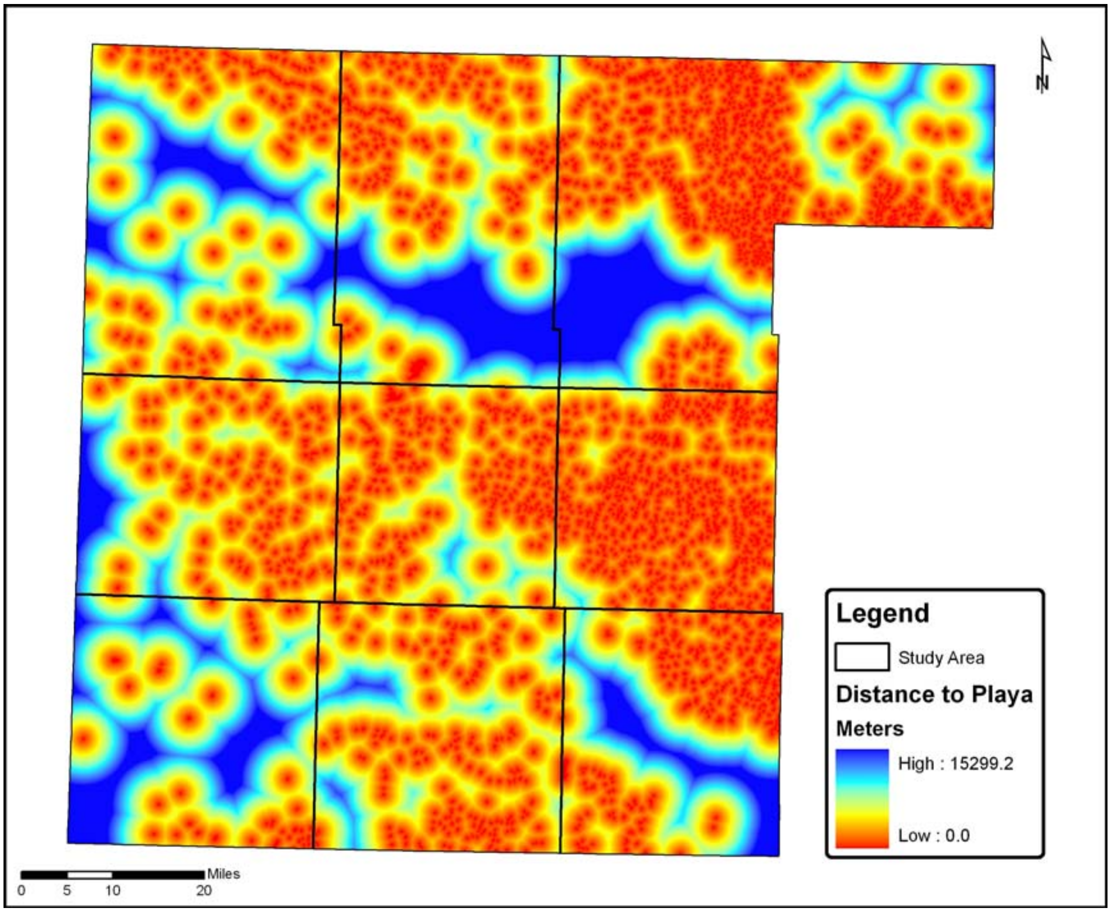


Figure 4.12: ‘Distance to Playa Lake’ variable generated from the SSURGO-extracted playa lakes.

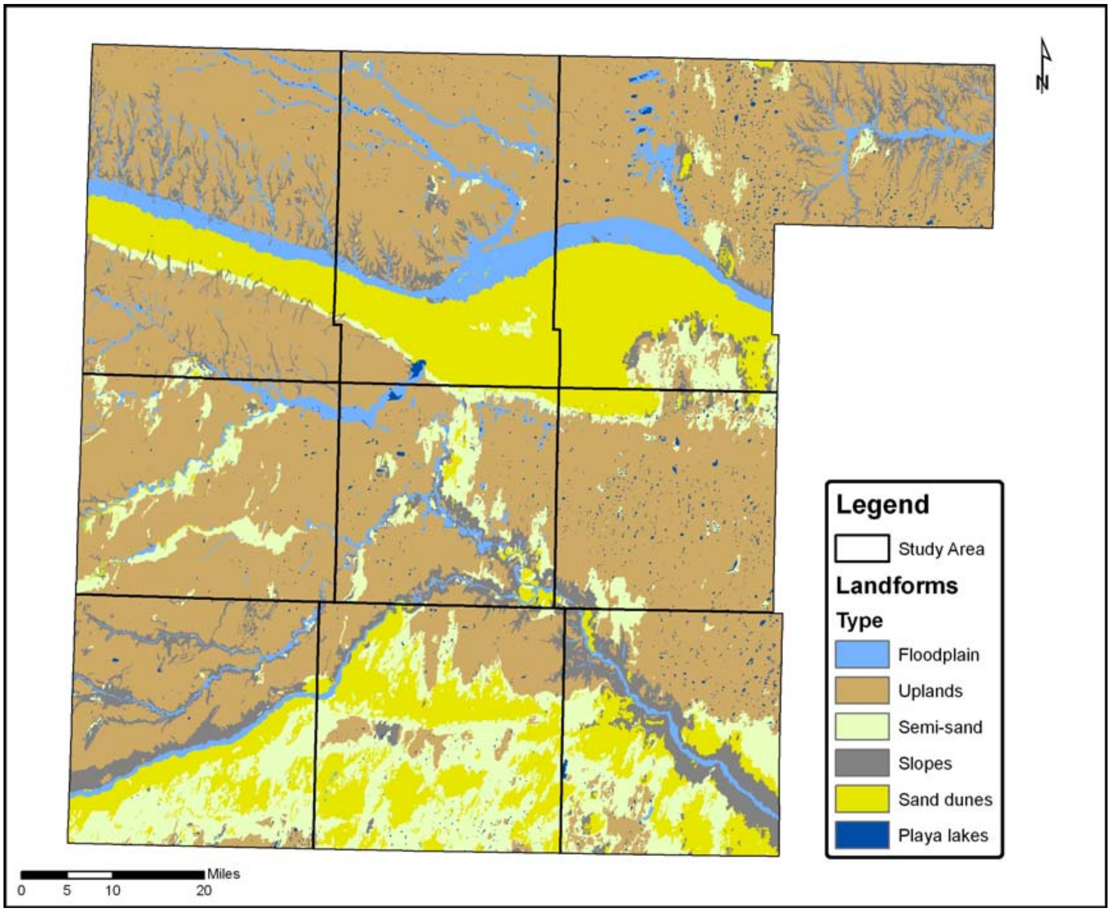


Figure 4.13: Landform map of the study area. Dataset created by generalizing SSURGO soils data; over 20,000 individual soil polygons were condensed into six potential landform categories.

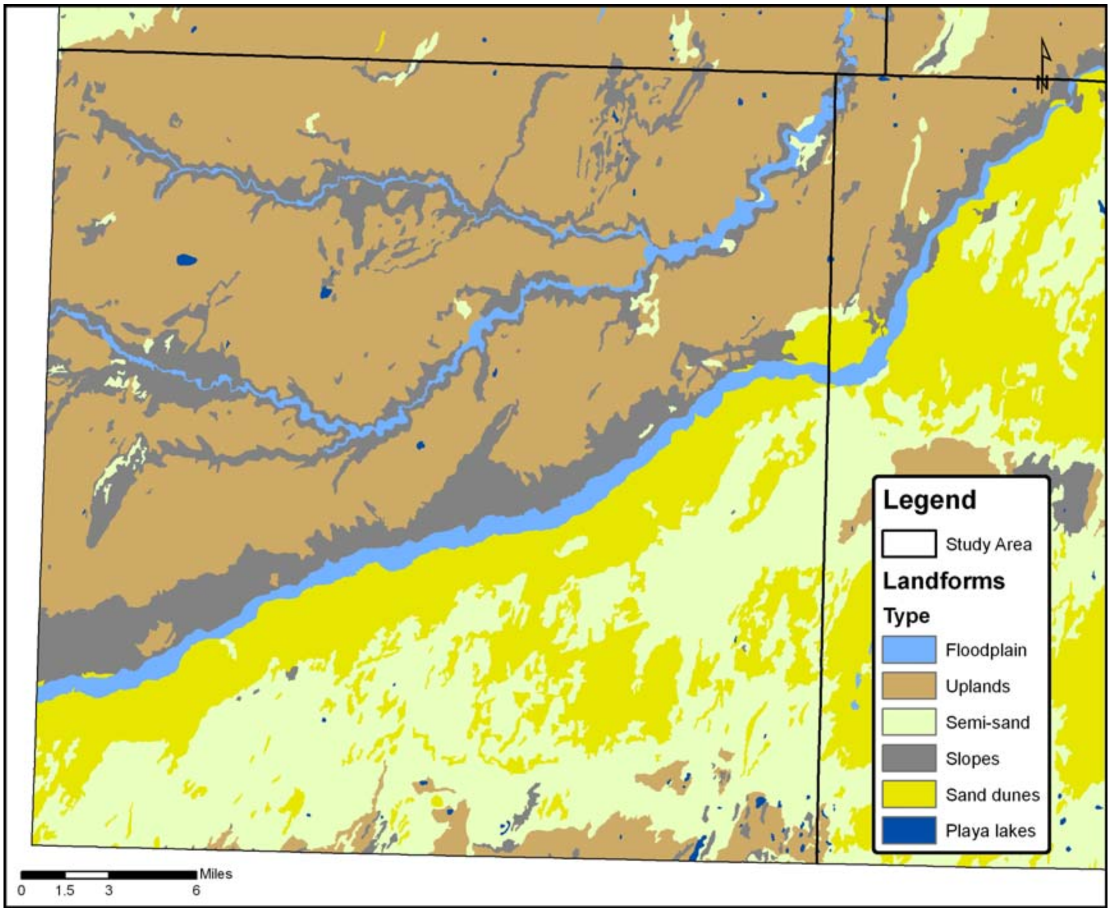


Figure 4.14: Landform map of Morton County. Larger cartographic scale displays the detail of the SSURGO source data (1:24,000).

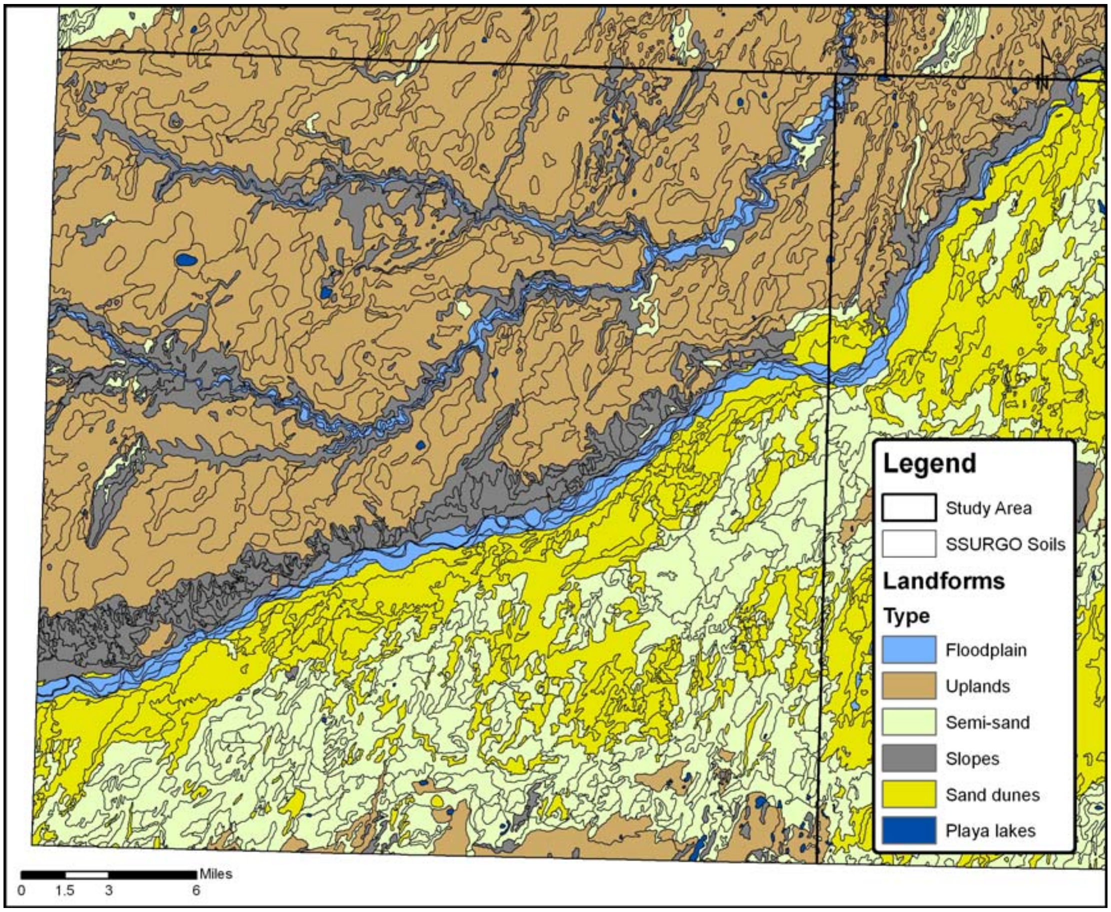


Figure 4.15: Landform map of Morton County (Fig 4.14), except with the SSURGO soil polygons displayed on top of the landform classifications. Displays the extent of landscape generalization possible from SSURGO data and the MLRA organizational concept.

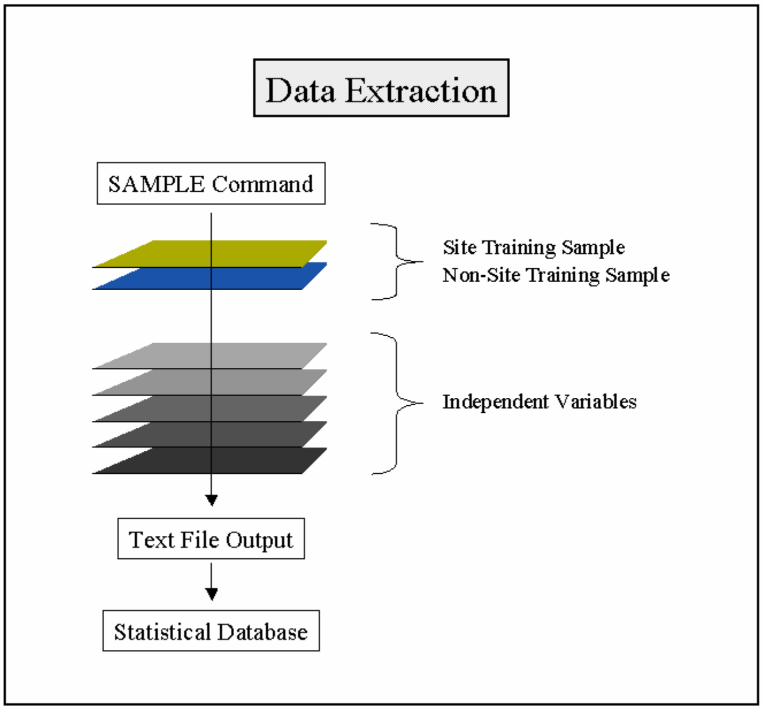


Figure 4.16: Data extraction using the 'Sample' Arc/INFO command for the Site Training and Non-Site Training locations. Output of the 'Sample' command is a text file that was imported to SPSS for analysis. 'Sample' command also used to extract data used in model testing.

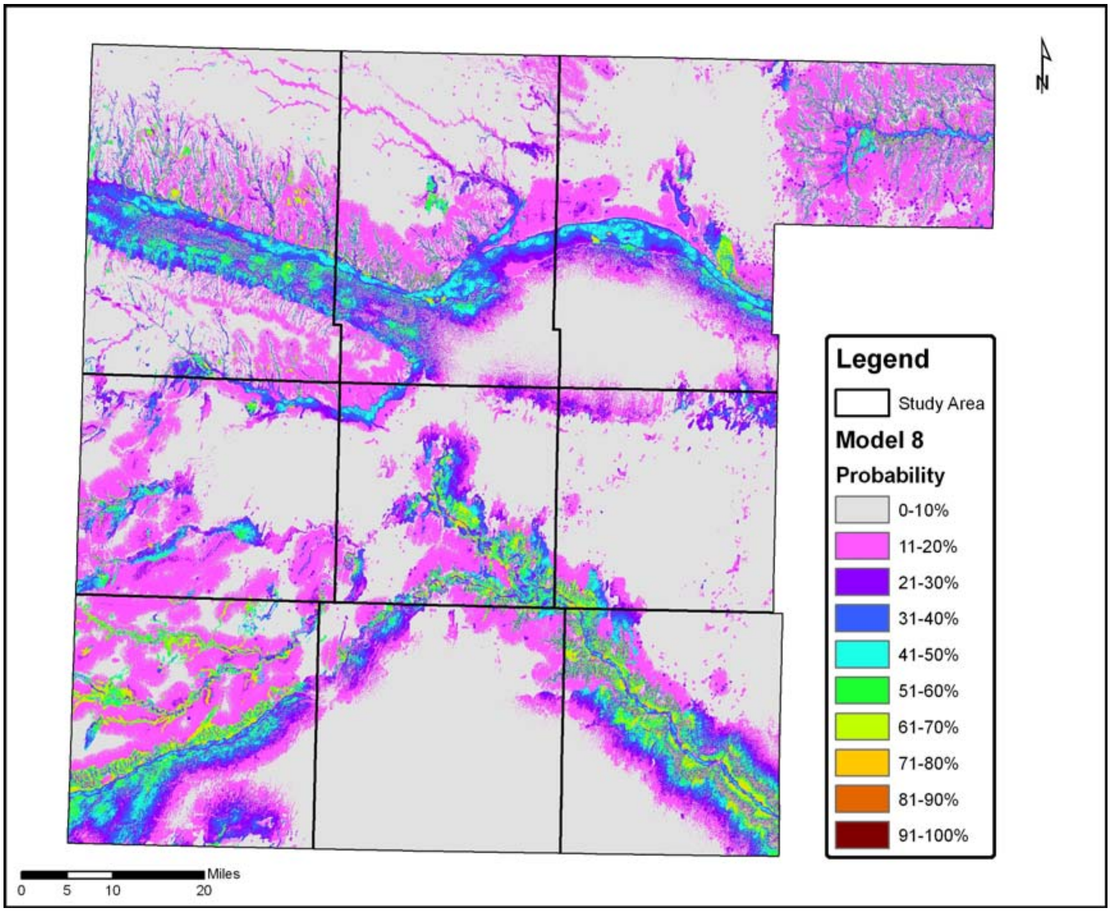


Figure 4.17: Binary logistic regression output of the archaeological predictive model. Each of the over 20 million land parcels in the study area were independently evaluated to produce a unique probability score.

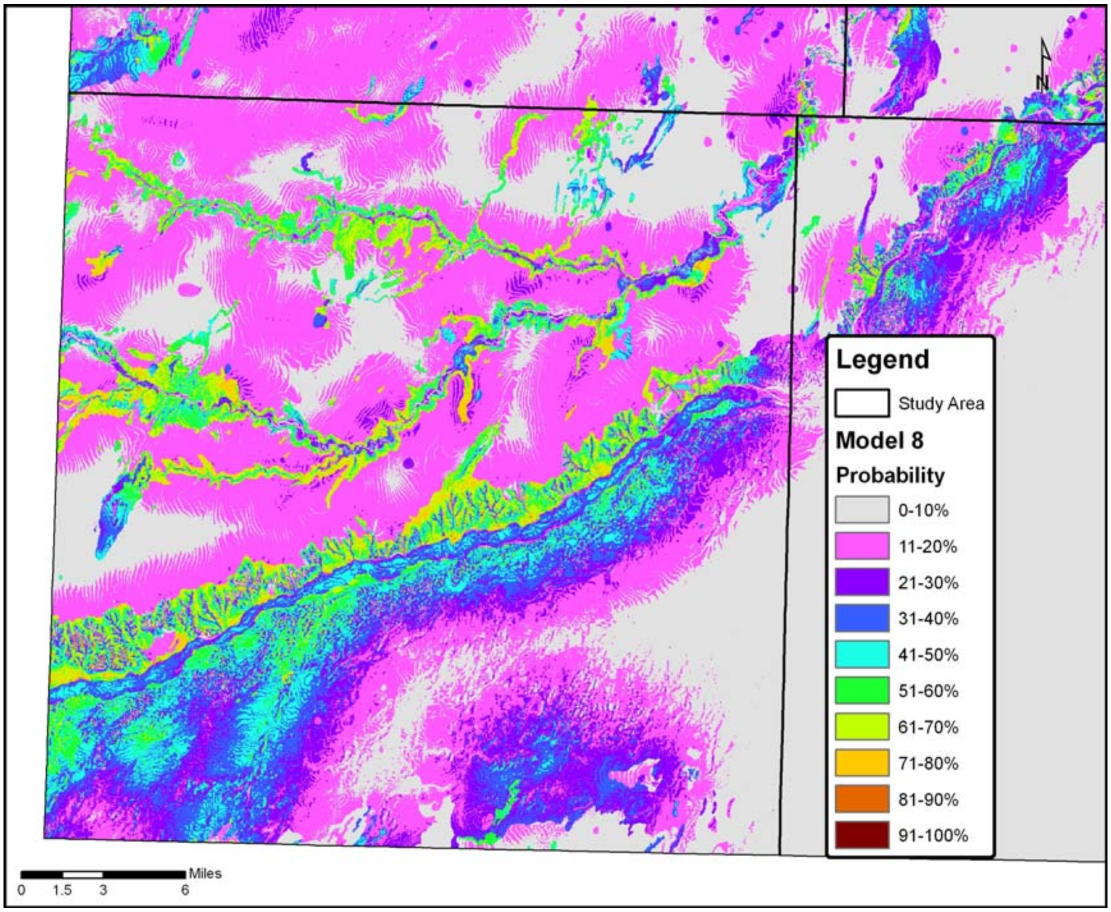


Figure 4.18: APM results in Morton County. Note the detail that becomes apparent as the cartographic scale increases.

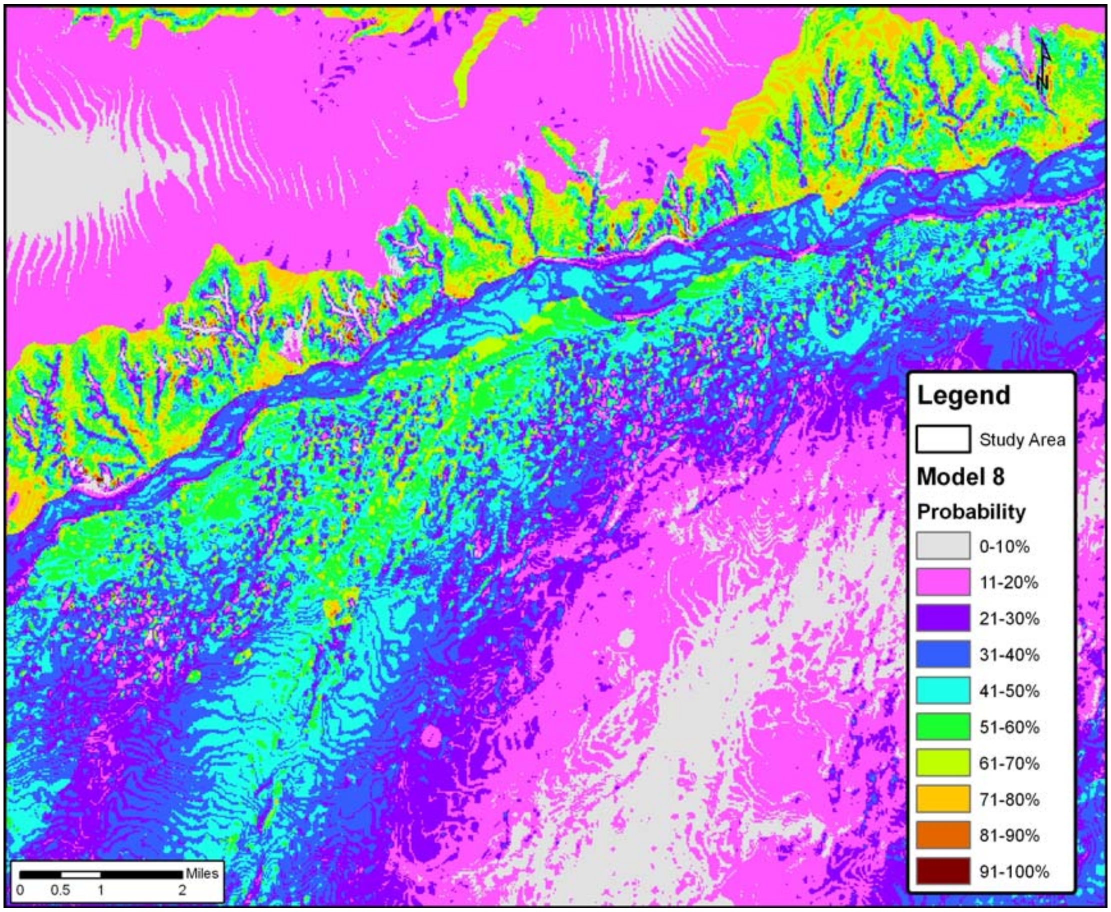


Figure 4.19: APM results in Morton County. Note the high variability of probability scores that become apparent as the cartographic scale increases. Viewing the model output at a coarse cartographic scale often hides the true distribution of probability scores.

Chapter 5

Model Testing

Model accuracy was assessed using the techniques described in Kvamme (1992). Specifically, the methods and logic for the accuracy assessment are reported below. The optimal modeling goal is to maximize the percentage of correctly identified land parcels in the site-present archaeological event class {S} and simultaneously minimize the total number of land parcels predicted in the site-present class {M}. The techniques used to meet this goal are a critical component of model testing.

Accuracy of the predictive model was measured primarily in terms of its ability to correctly classify both known site locations (S) and known non-sites (S'). A complete representation of model accuracy includes both the percentage of correctly identified sites and percentage of correctly identified non-sites. The percentage correct of sites represents the percentage of sites (S) that are correctly classified within the site-present class of the model (M), while the percentage correct of non-sites (S') represents the percentage of the site-absent class (S') correctly classified in the site-absent class of the model (M'). These two measures can be described as Pr(M|S) and Pr(M'|S'). Additional assessment metrics include the probability of a site occurring when the model predicts a site, Pr(S|M), and the probability of a site occurring when the model does not predict a site, Pr(S|M').

Kvamme (1988) indicates that to be considered useful, a predictive model must perform better than a random chance model. Using the metrics described above, and the base-rate probabilities, the model can be evaluated in a quantitative and defendable manner. Comparing the measures of model accuracy with the base-rate probabilities provides a method of quantifying model accuracy as a percentage increase over random. Computation of the random chance or base-rate models is discussed below.

Ultimately the goal of the model is to classify each land parcel into one of the two event classes (M and M'), yet the output of the regression is a probability score ranging between the values 0 – 1. In order to translate the continuous probability score into the binary classification of the event classes, a 'cut-point' in the range of probabilities must be established. For example, the standard cut-point is 0.5, meaning that any land parcel with a probability score of 0.5 or greater would be assigned to the site-present (M) class and any score less than 0.5 would be in the site-absent (M') class. This relationship is described mathematically as:

$$M = L \geq 0.5$$

and

$$M' = L \leq 0.5$$

where, L is the probability score of a given land parcel. Although 0.5 is the standard cut-point, the value can be shifted higher or lower based on modeling needs. Consider if the cut-point were moved, or ‘slid down’, to 0.4, the percentage of archaeological locations correctly identified would increase, but an associated decrease would occur in the percentage of non-site locations correctly identified. Increases in the site-present prediction accuracy are due to a larger land area being included in the site-present class (M) as the cut-point is lowered. Using this logic to the extreme, it is possible to correctly identify 100% of the archaeological sites by moving the cut-point to an extremely low number (0.01), however the model would accurately predict 0% of the non-site locations, and the site-present class (M) would occupy 100% of the landscape. This would offer no utility to land use managers as the screening component of the model would be useless. The relationship between probability score and percentage correct can be graphed for both the site-present and site-absent classes. An inverse relationship exists between the site-present and site-absent classes, meaning that as the percentage correct of the site-present class increases, an associated decrease in the percentage correct site-absent class occurs.

Three methods for determining the appropriate cut-point are found in the literature. First, the most basic approach is to use a 0.5 cut-point. This method is implemented by default in SPSS and represents the simplest conceptual approach. The second method utilizes a graphical interpretation of the cumulative correctly classified curves for both the site-present and site-absent classes. The cut-point is placed at the graphical intersection of the percentage correct lines for both the site-present and site-absent classes. The intersection cut-point represents the model optimum, that is, the cut-point in which the greatest percentage of site-present parcels and site-absent parcels are correctly classified simultaneously (Warren 1990a). An example of this type of curve is displayed in Figure 5.3. Third, Kvamme (1992) indicates that a predictive model should correctly identify at least 85% of the site-present sample; therefore the cut-point is established by determining the probability value at which 85% of the sites are correctly classified. The Mn/Model goes a step further in requiring their Phase 3 models correctly identify 85% of the sites and that the landscape area classified as site-present (M) does not occupy more that 33% of the total landscape (Hudak et al., 2000).

For the purposes of this model, the cut-point is set at the level in which 85% of the sites are accurately classified. The argument for the 85% classification accuracy stems from the rare nature of archaeological sites and the belief that it is more effective to lower the accuracy of the site-absent class than it is to not predict the location of a site (Kvamme 1992). Although accuracy of the site-present class should be maximized, it offers little utility if the land area predicted as site-present is 100% of the landscape. Limiting the maximum amount of land parcels classified as site-present is a way to place additional, functional constraints on the modeling process. While not explicitly adhered to in this report, the percent of landscape metrics are reported along with the 85%-derived cut-point.

Base-Rate Probabilities

A fundamental requirement of APM assessment is computation of the base-rate, or random chance, probabilities. A total of 226 sites are located within the study area, these sites occupy a total of 11,261 30m x 30m land parcels. The entire study area occupies 20,440,315 land parcels. Therefore, the base-rate or *apriori* probability of the site-present {S} event class can be calculated as:

Pr(S) = 11,261/20,440,315 = 000550

(0.05% of all land parcels contain a site)

And the site-absent class {S'} as:

Pr(S') = 20,429,054/20,440,315 =0.999449

(99.95% of all land parcels do not contain a site)

The event classes are mutually exclusive and represent all possible outcomes, *i.e.*, Pr(S) + Pr(S') = 1. The base-rate probabilities provide "pure-chance" probabilities for each archaeological event class. Using an example from Kvamme (1992), the "pure-chance" probabilities are analogous to the probability of identifying a site by throwing darts at a map. By chance, 0.05% of the darts would land on a site parcel and 99.95% would not. Establishing the base-rate probabilities for the two event classes sets the standard by which the predictive model is evaluated. In order to be considered effective, the model must “predict an event occurrence with probability greater than the event’s base-rate chance of occurrence” (Kvamme, 1992:28). Written mathematically, the previous statement is expressed as:

$$\Pr(S|M) > \Pr(S)$$

Where Pr(S|M) is the probability of a site given that the model predicts a site. The mathematical expression is the quantitative version of the statement that a model must perform better than random chance.

The calculated value of Pr(S) for the study area is artificially lower than reality due to the paucity of known archaeological sites in the majority of the study area. In order to better judge the true condition of the archaeological distribution is the study area, Pr(S) has also been calculated for Morton County. The extensive cultural survey of Morton County has resulted in the majority of known sites in the study area. Pr(S) for Morton County equals 0.003, indicating 0.3% of land parcels contain cultural material. While still very low, this value is nearly 10 times greater than the study area as a whole. Considering the Point of Rocks and Middle Spring area represents a heavily utilized location, the Pr(S) for Morton County is near the upper end of a plausible range of base-rate site-present probabilities.

Assessment of Training Data

In order to measure the effectiveness of the regression model, the output probability surface was reclassified into 10 groups of equal interval between 0 –1. The reclassified values for each site-present and site-absent land parcel were extracted from the GIS and graphed using Microsoft Excel. If the model is performing well, land parcels that contain an archaeological site should receive a high probability, while the site-absent class should receive a low probability. Graphically the two classes should appear as distinct clusters at either end of the probability range. Figures 5.1 and 5.2 display the histograms for each class.

The site-absent class (Figure 5.1) is classified very well, with 80% of the land parcels scoring below 0.2 and 86% scoring below 0.3. However, the site-present class (Figure 5.2) does not perform as well. Only a small proportion of the site-present class (3%) received a value above 0.8 and nearly 20% scored below 0.2. The remaining 77% of samples are classified between 0.21 and 0.80. The optimal distribution would be the mirror image of the site-absent class.

The graphical cut-point of the training data displays the cumulative distribution of all sample land parcels in both event classes (Figure 5.3). No reclassification of the data into classes was required to create this graph. By including all the data points the output curve is much smoother and provides a better visual depiction of the model accuracy. The graphical cut-point for the training data is found at the intersection of the two curves, in this case at the probability score equal to 0.40. At this level, the model is accurately predicting approximately 82% of both the site-present and site-absent classes. 85% of the site-present class is accurately predicted at a cut-point of 0.36. However, as stated earlier an upward bias in the predictive power of the model is expected when using the training data to evaluate performance.

Assessment of Testing Data

Assessment of the testing data follows the same procedure as the training data; the model was reclassified into 10 classes to construct the histograms, while the raw data were used to create the cut-point graph. Figures 5.4 and 5.5 display the results of the testing data assessment. The site-absent class again performed very well with 80% of the sample receiving a score below 0.2 and 87% scoring below 0.3. Similarity of the results between the training and testing samples is expected due to the random distribution of the site-absent sample.

Results of the site-present class are not a good as the training sample, as would be expected. However the reduction in power represents a significant decrease. 32% of the sample scored below 0.2 and only 1.5% scored above 0.8. The remaining 66% scored between 0.21 and 0.80. Decreased power of the model to distinguish between the site-present and site-absent testing samples is indicated by the large number of sites in the 0.1-0.2 probability class. Graphical determination of the cut-point for the testing data is approximately 0.16 (Figure 5.6). At this level, approximately 75% of both the site-present and site-absent classes are correctly classified. In order to correctly classify 85% of the site-present sample, the cut-point must be slid down to

0.11. This represents a significant reduction from the 0.36 cut-point of the training data.

A graphical cut-point was also created for the testing sample within Morton County (Figure 5.7). Results indicate the model is performing better in Morton County than in the study area as a whole. The graphical intersection is found at a 0.29 cut-point, in which 74% of the site-absent and site-present classes are correctly identified. A cut-point of 0.18 results in an 85% classification accuracy of the site-present class and 60% in the site-absent class. The increased accuracy within Morton County is encouraging for the model. Because of the heavy bias towards sites in Morton County and because this area represents the most complete archaeological survey, it is more reflective of the true potential distribution of archaeological sites in the study area. Additional Morton County metrics are reported below.

For the entire study area, 85% of the testing sample site data and 60% of the non-site testing data are correctly classified at the 0.11 cut-point. Additional assessment metrics can be derived by comparing the total amount of land parcels with the amount of site-present parcels classified within a particular probability class (Figure 5.8). If the curves of the known site samples have higher values than the overall landscape, the model is performing better than random chance. By manipulating the data used in the construction of Figure 5.8, it is possible to exactly quantify the model's percentage increase over random for any given probability class (Figure 5.9). Results indicate that at the 0.15 cut-point, the model is performing 42% better than random chance.

Graphical analysis indicates the 0.11 cut-point value represents approximately a 30% gain over a random classification (Figure 5.9). Dividing the study area into the site-present class (M) and the site-absent class (M') at the 0.11 cut-point, results in 41% of the study area assigned to the site-present class (M) and 59% assigned to the site-absent class (M') (Figure 5.10). The Phase 3 components of the Mn/Model required 85% prediction accuracy within 33% of the land area. Although the 33% value was not attained by the current model, this level of accuracy exceeds those reported in the beginning Phase 1 models created for the Mn/Model.

At the 0.11 cut-point, the probability levels associated with both event classes can be written using Kvamme's notation as:

$$\Pr(M|S) = 0.8498$$

$$\Pr(M|S') = 0.3956$$

where, Pr(M|S) is the probability that the model correctly identifies a site given that a site is actually present, and Pr(M|S') is the probability that the model correctly identifies a non-site given that a site is actually not present (Kvamme, 1992:33).

As stated earlier, for a predictive model to be considered successful, the probability of a site occurring given the model specifies a site, Pr(S|M), must be greater than the base-rate probability Pr(S), calculated at 0.00055. Pr(S|M) is the reverse conditional of Pr(M|S) and can be estimated using Bayes' Theorem:

$$\Pr(S|M) = \frac{\Pr(M|S)\Pr(S)}{\Pr(M|S)\Pr(S) + \Pr(M|S')\Pr(S')}$$

Using the values already determined, this equation yields:

$$\Pr(S|M) = \frac{(.8498)(.00055)}{(.8498)(.\ 00055) + (.3965)(.9994)}$$

$$\Pr(S|M) = 0.001182$$

$$\Pr(S|M) > \Pr(S)$$

$$0.001182 > 0.00055$$

Results indicate the probability of site occurring given the model predicts a site is equal to 0.001182; therefore Pr(S|M) is greater than Pr(S), establishing that the current model is more effective than random chance. Although Pr(S|M) is very low, it is due to the low base-rate probability and the fact that archaeological sites are rare on the landscape. Approximately 0.1% of the land parcels in the site-present area (M) will contain a site. Stated another way, if the model predicts a site, the probability of a site occurring is Pr(S|M) / Pr(S), or (.001182)/(.00055), = 2.15 times more likely than random chance alone. Considering that over 20 million land parcels are in the study area and over 8.2 million parcels assigned to the site-present class, this represents a significant gain over a random chance model (Kvamme 1992).

Using this same methodology it is possible to estimate the probability of a site occurring given the model predicts a non-site, or Pr(S|M'). Small changes to the above equation result in:

$$\Pr(S|M') = \frac{\Pr(M'|S)\,\Pr(S)}{\Pr(M'|S)\,\Pr(S) + \Pr(M'|S')\,\Pr(S')}$$

Using the values already determined, this equation yields:

$$\Pr(S|M') = \frac{(.1501)(.00055)}{(.1501)(.00055) + (.6043)(.9994)}$$

$$\Pr(S|M') = 0.000137$$

Calculation of Pr(S|M') indicates 0.013% of the land parcels in the site-absent area (M') will contain a site. The probability of finding a site in the area predicted as site-absent (M’) is a factor of 10x smaller that in the site-present area (M), Pr(S|M) = 0.001 vs. Pr(S|M’) = 0.0001. Compared to the base-rate probability, the probability of finding a site in the area predicted as site-absent (M) is calculated as Pr(S|M’) / Pr(S) or (.00013/.00055) = 0.24 times as likely as pure-chance. These values represent a significant decrease from the base-rate probability and indicate the model is effectively classifying the landscape into site-present and site-absent classes.

Morton County

Development of this predictive model was intended to utilize the large number of known sites in Morton County to guide additional archaeological investigation within the under-surveyed remainder of the study area. In that way, the spatial relationship determined from a well-known area could be used to find new sites in the less-known area. Morton County offered a good case study because it shares an environmental similarity with the rest of the study area. A true test of the model is its performance in predicting the sites in Morton County. Besides meeting the accuracy criteria established by Kvamme (1990), for the model to be considered a success, the results of the model in Morton County should outperform the results for the study area as a whole.

Using the same methodology and nomenclature, the results of the model in Morton County alone can be summarized as follows:

$$Pr(S) = 0.00304$$

$$Pr(S') = 0.99695$$

A cut-point of 0.18 will accurately predict 85% of known sites and 60% of non-sites in 39% of the land area. The probability of a site occurring in the area predicted as site-present is calculated as: $Pr(S|M) = 0.00629$. Comparison of the predicted site-present probability and the base-rate site-present probability indicates that $Pr(S|M)$ is greater than $Pr(S)$ ($0.00629 > 0.00304$), therefore the model is outperforming random chance alone. Due to the heavy bias towards Morton County in the site-present input data, it is encouraging that the model is performing better than the overall study area in terms of percentage correct at a higher cut-point.

Additional metrics for Morton County indicate the probability of finding a site in the area predicted as site-present is 2.07 times more likely than random chance alone. This also compares well with the similar metric derived for the entire study area (2.07 versus 2.15). Also, the probability of finding a site in the area predicted as site-absent is 0.25 times the base-rate probability. Considering the similarities in performance between Morton County and the study area as a whole, it appears that the initial concept of using the Morton County as a basis for predicting sites in the remainder of the study area was valid.

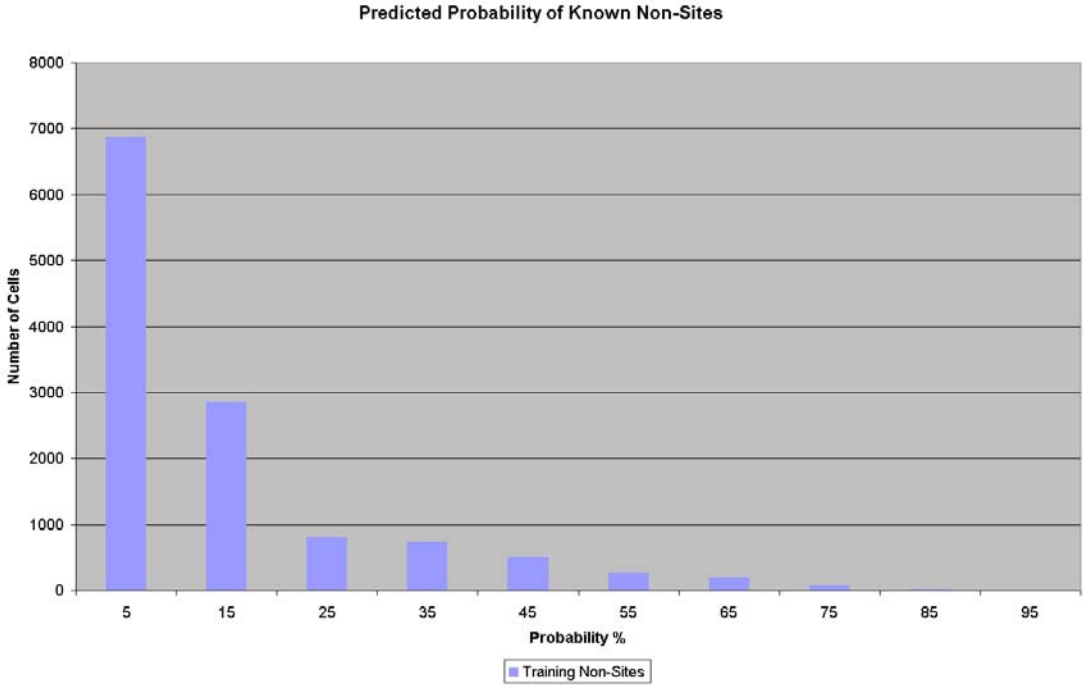


Figure 5.1: Histogram distribution of training site-absent land parcels.

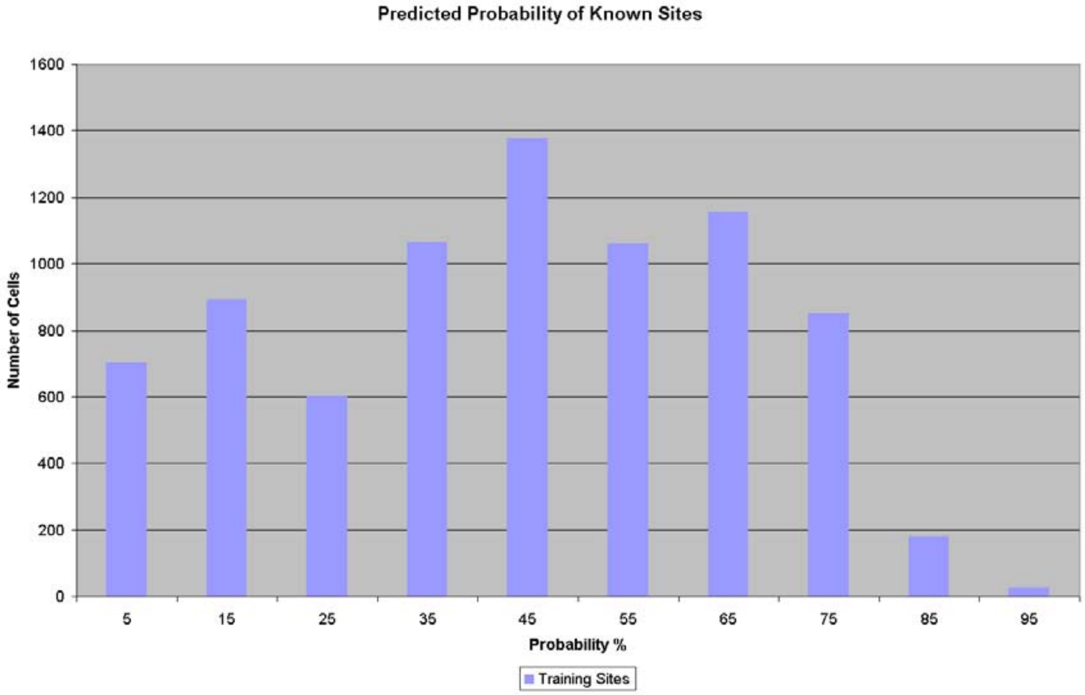


Figure 5.2: Histogram distribution of training site-present land parcels.

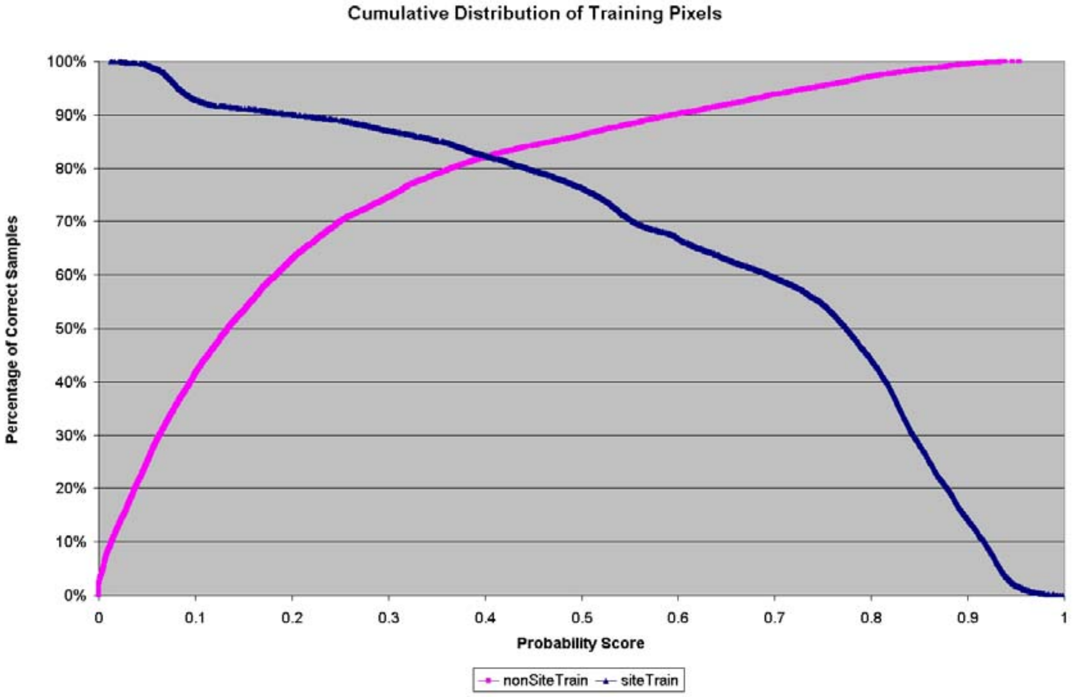


Figure 5.3: Graphical intersection of probability distributions of the site-absent and site-present training samples.

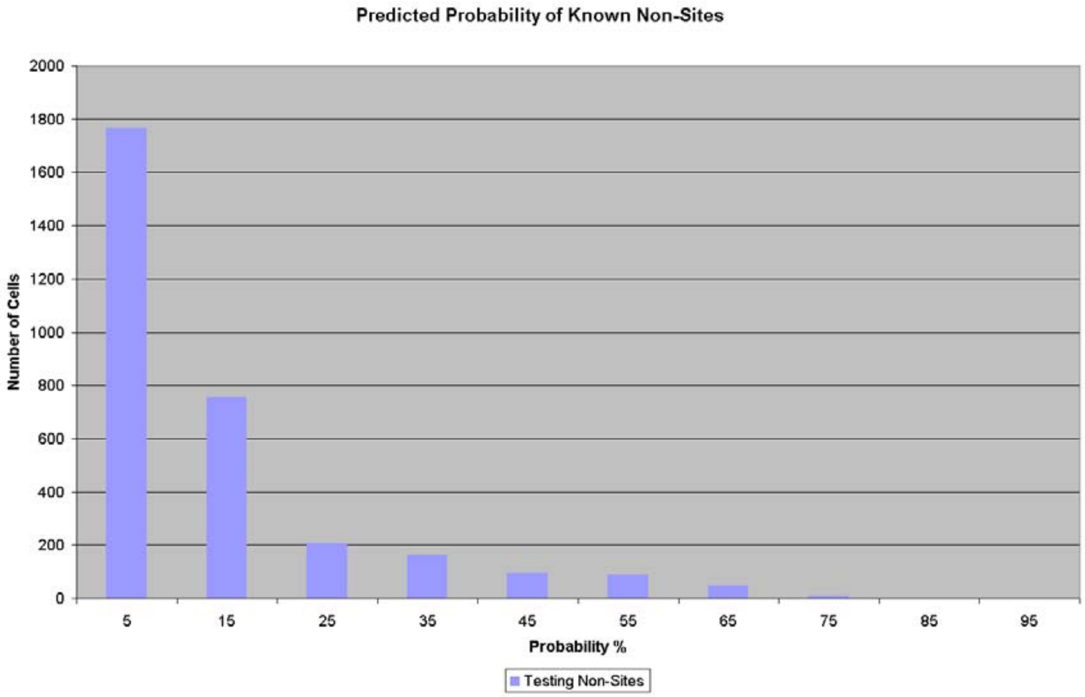


Figure 5.4: Histogram distribution of testing site-absent land parcels

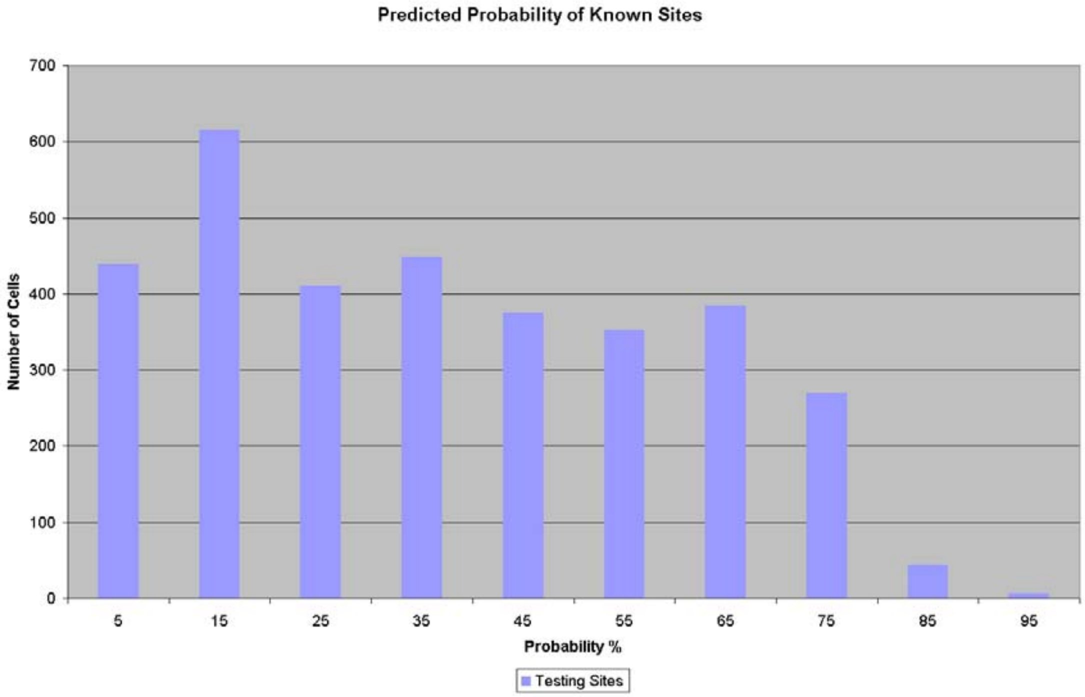


Figure 5.5: Histogram distribution of testing site-present land parcels

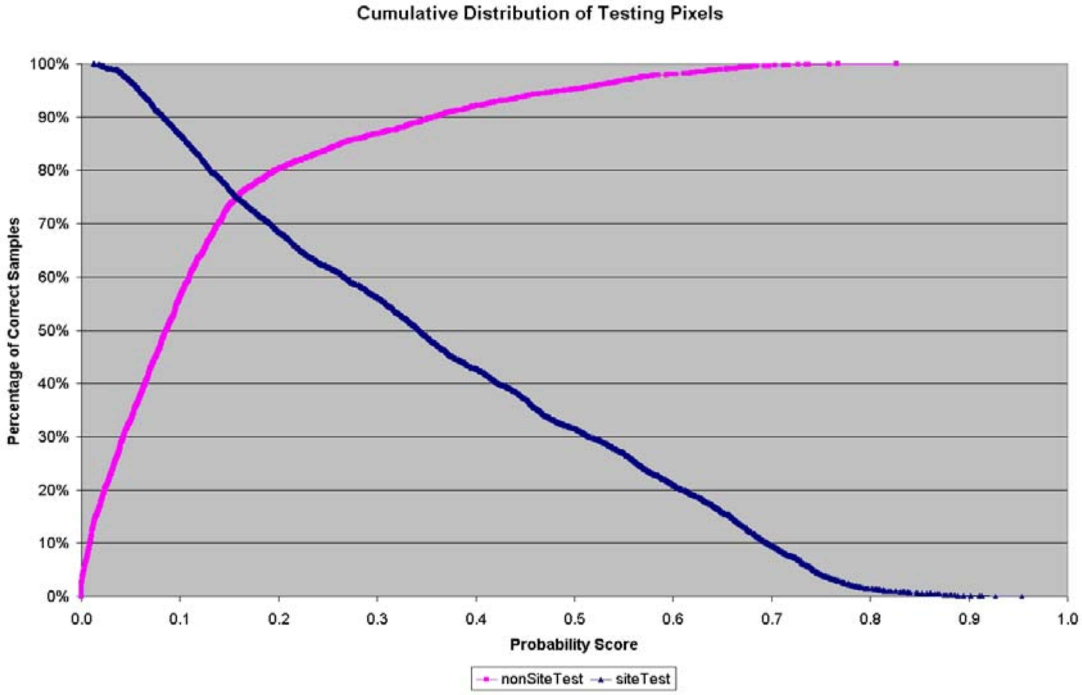


Figure 5.6: Graphical intersection of probability distributions of the site-absent and site-present testing samples for the entire study area.

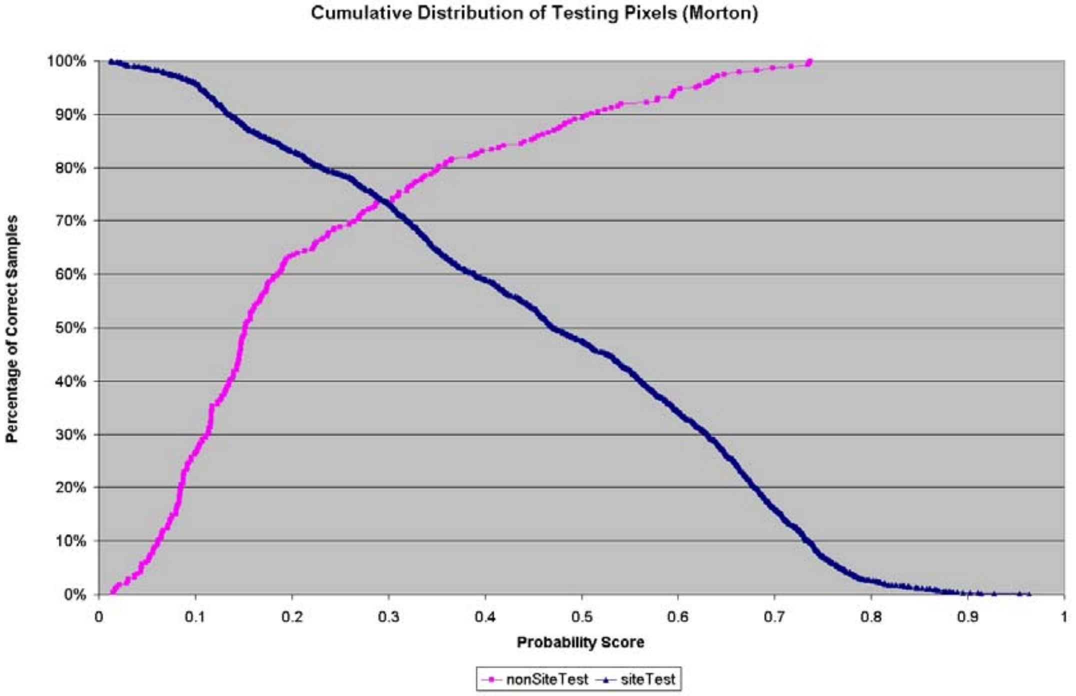


Figure 5.7: Graphical intersection of probability distributions of the site-absent and site-present testing samples in Morton County.

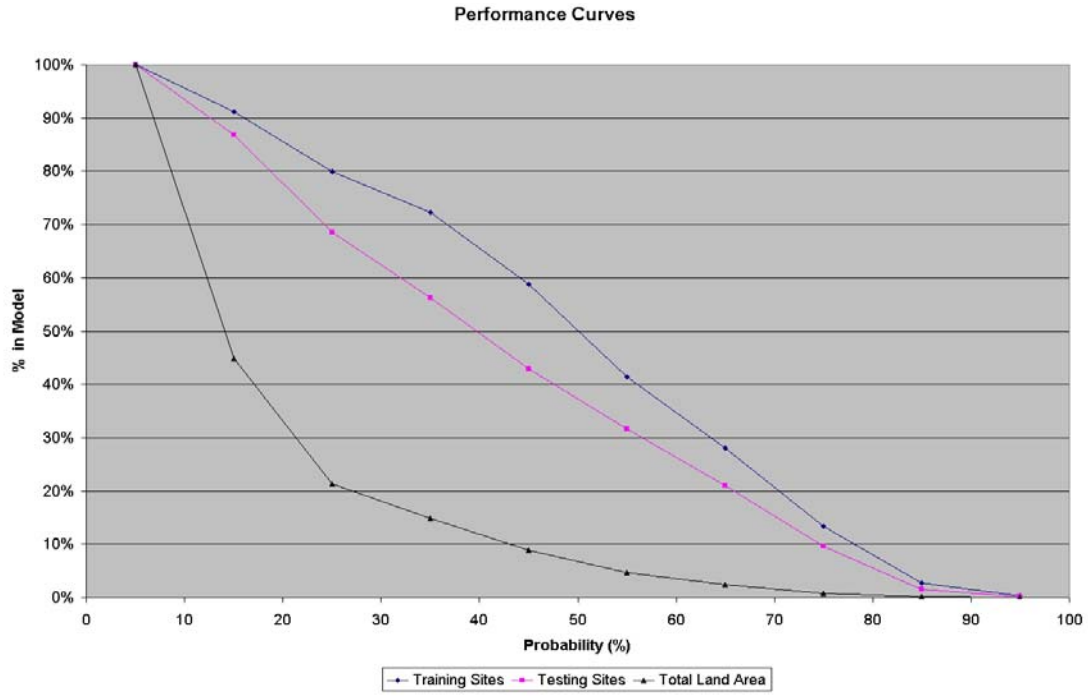


Figure 5.8: Distribution of the training and testing site-present samples in comparison to the distribution of probability scores for the landscape as a whole.

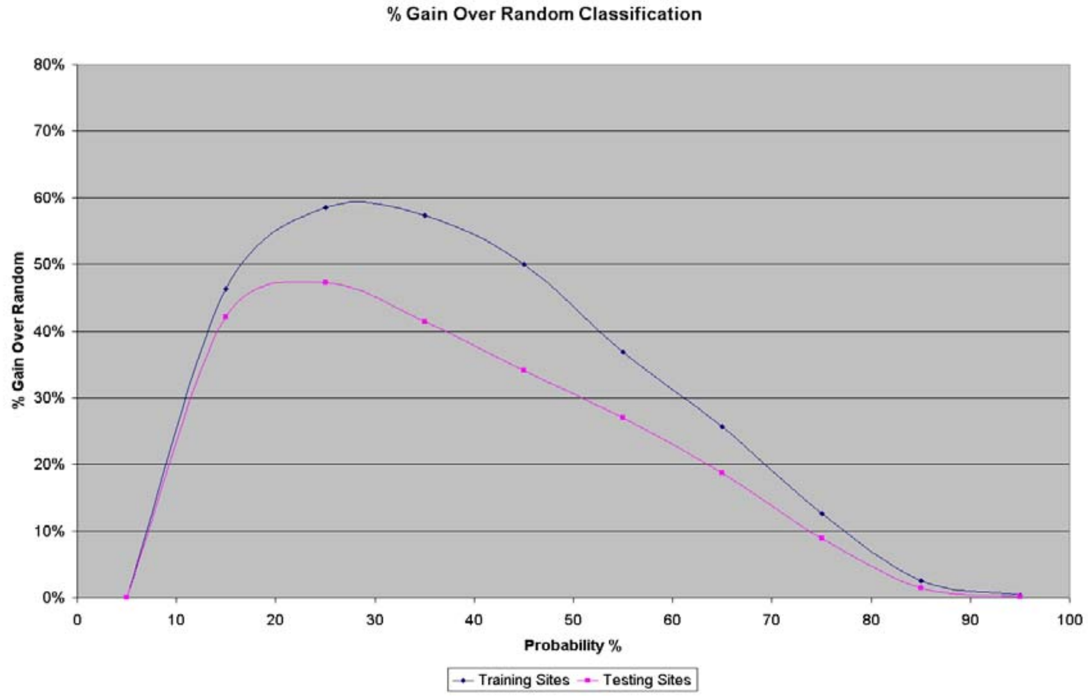


Figure 5.9: Normalized version of Figure 5.8. Displays the percentage increase over a random classification for any probability cut-point. For example, at a 0.15 cut-point the testing data is 42% better than a random classification.

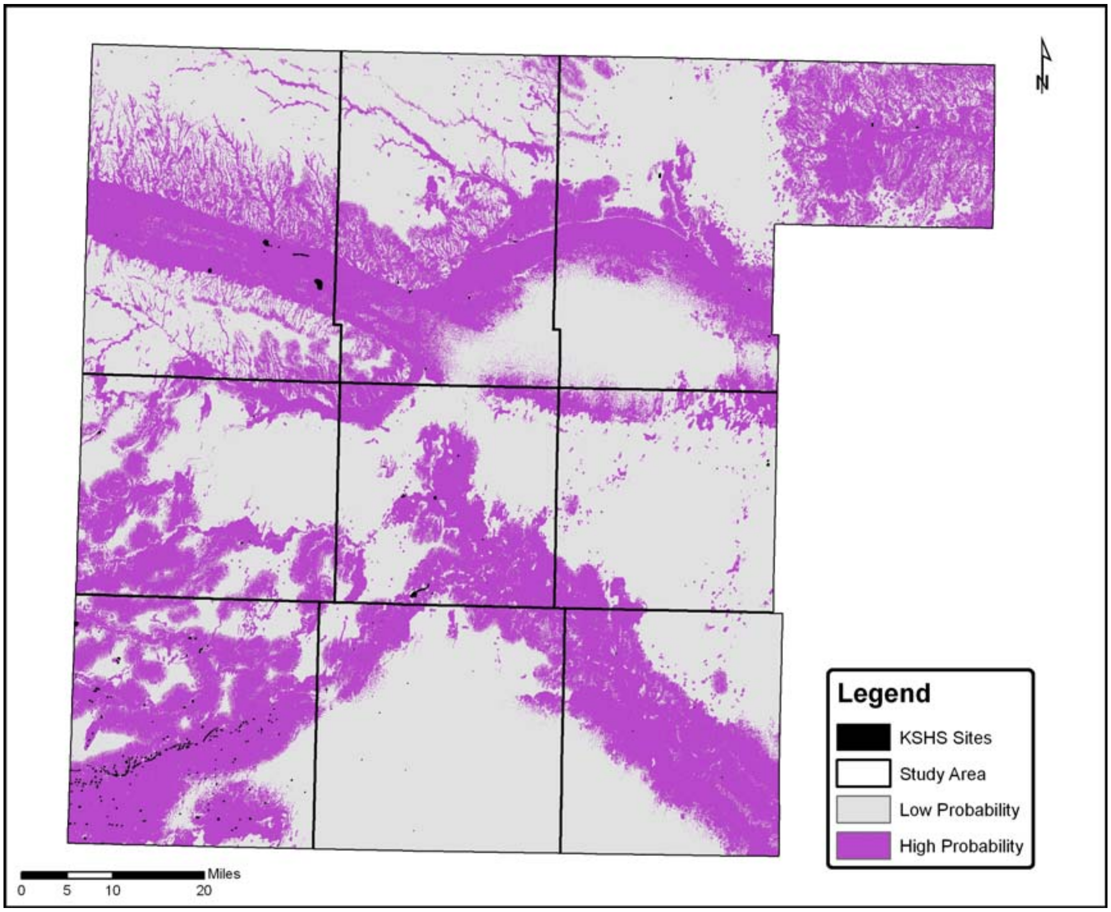


Figure 5.10: Binary classification of the landscape based upon a 0.11 cut-point. Land parcels with a probability score greater than 0.11 are coded as medium to high probability and areas less than 0.11 are coded low probability. In total, the area coded as medium and high probability represents 41% of the total landscape. Archaeological sites from the KSHS database are also shown for comparison.

Chapter 6

Conclusion

Results of the model assessment indicate the APM can be considered a successful model. Although the model is not as powerful as hoped, its results represent a significant increase from a random classification, thereby satisfying Kvamme's standard. Potential improvements to the modeling accuracy include increasing the number of sites included in the input data, and collecting new sites from field surveys designed to sample all landforms equally. Based on the majority of sites used to build the model, the model output identifies the high probability areas for discovering surficial or shallowly buried, Late-Prehistoric (last 2,000 yrs) cultural material. High-probability areas of the model occur in landscape positions close to water with a large degree of relief and that tend to be exposed. Areas along river drainages and around playa lakes score better than the large expanse of upland locations between the major rivers and in floodplains. Due to the lack of subsurface data, no conclusions about the location or extent of buried sites can reliably be drawn from this model. These results are in agreement with the conclusions put forth by Brown (1977).

This model represents an attempt to quantify the culture-landscape relationship for southwest Kansas. It is hoped that additional field surveys in the future will utilize this model, and the data used to create it, to aid in the development of stratified sampling designs. Designing field surveys with the creation of predictive models in mind would make the development of future models more methodologically sound and likely more powerful. In many ways the conclusions drawn from this model are not surprising, however the utility of being able to visualize how the combination of geographic factors are distributed on the landscape should not be underestimated.

In light of the critiques of archaeological predictive modeling it is important to note that the knowledge of the archaeological condition of this area is extremely low. It is the author's opinion that any tool that can help to increase the amount of sites for analysis represents a significant step forward for the region. Model conclusions about the potential distribution of additional cultural resources are not intended as a theoretical explanation of why the sites are there. Instead this model is meant to be a guide for the formulation of new systematic surveys of the area so that those of a theoretical bent can have a significant amount of material to construct holistic interpretations of site locations in the region.

Modeling techniques utilized in this study are scalable and could be adapted for a statewide approach. MLRA designations facilitate the partitioning of the landscape into model-ready subsets and the ongoing development of the KSHS GIS database provide the fundamental components required for further model development. Considering the recent development of statewide predictive models in Minnesota, North Carolina, and Vermont, Kansas is well positioned in terms of data access and archaeological content to utilize this methodology for a statewide model.

The role of Geographic Information Systems (GIS) in the development of this model cannot be underestimated. From an analytical perspective, consider the amount of calculations required to develop this model. The study area occupies over 20 million land parcels. For each of these land parcels, nine variables were created, resulting in a total of more than 180 million values for the environmental variables alone. In addition, the number of archaeological site locations, non-site locations, and the number of calculations required to compute the statistical model must also be considered. The amount of calculations required to develop the model could not be completed without GIS software and modern computing power. Besides the analytical and data storage aspects, the visualization component of the GIS provides a fundamental capacity to understanding the complex spatial patterns inherent to a dataset of this type.

Ultimately the quantitative approach to archaeological predictive modeling used for this model was effective. Results of this model provide a basis for further investigation of the High Plains region of southwest Kansas, an area in need of more exploration by archaeologists. One of the primary motivations for undertaking this process was the extensive amount of cultural materials in the hands of local collectors. These individuals have spent a lifetime gathering material from this area that has not been incorporated into the collective body of academic knowledge. It is not that archaeological materials do not exist in the region, they just have not been fully explored. If the archaeological community wants to understand the extent of cultural material, a comprehensive approach involving quantitative modeling, field surveys, and human interviews will be required. The model discussed in this report will serve as a starting point for the exploration of this archaeologically valuable area.

References Cited

Allen, K. M. S., S. W. Green, and E. B. W. Zubrow, eds. 1990. *Interpreting Space: GIS and Archaeology*. Edited by D. F. Marble and D. J. Peuquet, *Applications of Geographic Information Systems*. London: Taylor & Francis.

ArcNews. 2006. Modeling Archaeological Sensitivity in Vermont with GIS. *ArcNews*, 2006.

Bamforth, D. B. 1988. *Ecology and Human Organization of the Great Plains*. Edited by M. Jochim, *Interdisciplinary Contributions to Archaeology*. New York: Plenum Press.

Bartlett, R. B., L. C. Bement, and R. L. Brooks. 1993. A Cultural Resource Assessment of Promontories in Western Oklahoma, 42. Norman: The University of Oklahoma, Oklahoma Archeological Survey.

Brosowske, S. D., and L. C. Bement. 1998. Pedestrian Survey of Playa Lake Environments in Beaver and Texas Counties, Oklahoma, 61. Norman: Oklahoma Archaeological Survey.

Brown, K. L. 1977. Late Prehistoric Settlement Patterns in Southwestern Kansas: A Model. Masters Thesis, Anthropology, University of Kansas, Lawrence.

Burns, D. 2001. Discussion on southwestern Kansas archaeology.

Butzer, K. W. 1982. *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge: Cambridge University Press.

Campbell, J. S., and W. C. Johnson. 2004. Temporal Predictive Model for Fort Hood, Texas: A Pilot Study in the Cowhouse Creek Drainage, 47. Fort Hood: United States Army.

Carr, C., ed. 1985. *For Concordance in Archaeological Analysis: Bridging Data Structure, Quantitative Technique and Theory*. Kansas City: Westport Publishers.

Church, T., R. J. Brandon, and G. R. Burgett. 2000. GIS Applications in Archaeology: Method in Search of Theory. In *Practical Applicaitons of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Wescott and R. J. Brandon, 135-156. London: Taylor & Francis.

Clark, W. A. V., and P. L. Hosking. 1986. *Statistical Methods for Geographers*. New York: John Wiley & Sons.

Dalla Bona, L. 2000. Protecting Cultural Resources through Forest Management Planning in Ontario Using Archaeological Predictive Modeling. In *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Westcott and R. J. Brandon, 73-99. London: Taylor & Francis.

Dobson, J. E., E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley. 2000. LandScan: a Global Population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing* 66 (7):849.

Duncan, R. B., and K. A. Beckman. 2000. The Application of GIS Predictive Site Location Models within Pennslyvania and West Virginia. In *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Wescott and R. J. Brandon, 33-58. London: Taylor & Francis.

Ebert, J. I. 2000. The State of the Art in "Inductive" Predictive Modeling: Seven Big Mistakes (and Lots of Smaller Ones). In *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Wescott and R. J. Brandon, 129-134. London: Taylor & Francis.

Gaffney, V., and P. M. v. Leusen. 1995. Postscript - GIS, Environmental Determinism and Archaeology. In *Archaeology and Geographical Information Systems: A European Perspective*, eds. G. Lock and Z. Stancic, 367-382. London: Taylor & Francis.

Gibbon, G. 2000. Appendix A: Archaeological Predictive Modeling: An Overview. In *A Predictive Model of Precontact Archaeologial Site Location for the State of Minnesota*, eds. G. J. Hudak, E. Hobbs, A. Brooks, C. A. Sersland and C. Phillips. St. Paul: Minnesota Department of Transportation.

Hofman, J. L., ed. 1996. *Archeology and Paleoecology of the Central Great Plains*. Vol. 48, *Arkansas Archaeological Survey Research Series*. Fayetteville: Arkansas Archeological Survey.

Hofman, J. L., and R. W. Graham. 1998. The Paleo-Indian Cultures of the Great Plains. In *Archaeology on the Great Plains*, ed. W. R. Wood, 87-139. Lawrence: University Press of Kansas.

Hofman, J. L., B. Logan, and M. J. Adair. 1996. Prehistoric Adaptation Types and Research Problems. In *Archaeology and Paleoecology of the Central Great Plains*, ed. J. L. Hofman, 203-220. Fayetteville: Arkansas Archaeological Survey.

Holliday, V. T. 1997. *Paleoindian: Geoarchaeology of the Southern High Plains*. Austin: University of Texas Press.

———. 2000. The Evolution of Paleoindian Geochronology and Typology on the Great Plains. *Geoarchaeology: An International Journal* 15 (3):227-290.

Hudak, G. J., E. Hobbs, A. Brooks, C. A. Sersland, and C. Phillips. 2000. A Predictive Model of Precontact Archaeological Site Location for The State of Minnesota. St. Paul: Minnesota Department of Transportation.

Jochim, M. A. 1976. *Hunter-Gather Subsitence and Settlement: A Predictive Model*. Edited by S. Struever, *Studies in Archaeology*. New York: Academic Press.

Johnson, W.C. and J.S. Campbell. 2004. Playa Lake GIS Database. Lawrence: Data Access and Support Center

Johnson, W. C., and K. Park. 1996. Late Wisconsian and Holocene Environmental History. In *Archaeology and Paleoecology of the Central Great Plains*, ed. J. L. Hofman, 3-28. Fayetteville: Arkansas Archaeological Survey.

Judge, W. J., and L. Sebastian, eds. 1988. *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*. Denver: U.S. Bureau of Land Management, Department of Interior.

Kansas Geological Survey. 1984. Physiographic Map of Kansas. Lawrence: Kansas Geological Survey.

Kansas State Historical Society. 2002. Official Kansas Registry of Archaeological Sites. Topeka.

Krist, F. J. 2001. A predictive model of Paleo-Indian subsistence and settlement. *DAI* 62 (07A):381.

Kuna, M. 2000. Session 3 discussion: Comments on archaeological prediction. In *Beyond the Map: Archaeology and Spatial Technologies*, ed. G. Lock, 180-186. Amsterdam: IOS Press.

Kvamme, K. L. 1988. Development and Testing of Quantitative Models. In *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*, eds. W. J. Judge and L. Sebastian, 325-428. Washington, D.C.: U.S. Government Printing Office.

———. 1990. The Fundamental Principles and Practice of Predictive Archaeological Modeling. In *Mathematics and Information Science in Archaeology: A Flexible Framework*, ed. A. Voorrips, 257-295. Bonn: Holos.

———. 1992. A Predictive Site Location Model on the High Plains: An Example with an Independent Test. *Plains Anthropologist* 37:19-40.

Kvamme, K. L., and T. A. Kohler. 1988. Geographic Information Systems: Technical Aids for Data Collection, Analysis and Display. In *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*, eds. W. J. Judge and L. Sebastian, 493-548. Washington, D.C.: U.S. Government Printing Office.

Lang, N. 2000. Beyond the Map: Harmonising Research and Cultural Resource Management. In *Beyond the Map: Archaeology and Spatial Technologies*, ed. G. Lock, 214-228. Amsterdam: IOS Press.

Lock, G., ed. 2000. *Beyond the Map: Archaeology and Spatial Technologies*. Vol. 321, *NATO Science Series A: Life Sciences*. Amsterdam: IOS Press

Lock, G., and Z. Stancic, eds. 1995. *Archaeology and Geographical Information Systems: A European Perspective*. London: Taylor & Francis.

Oklahoma Archeological Survey. 2002. Offical Oklahoma Registry of Archaological Sites. Norman.

Parker, S. 1985. Predictive Modelling of Site Settlement Systems Using Multivariate Logistics. In *For Concordance in Archaeological Analysis: Bridging Data Structure, Quantitative Technique, and Theory*, ed. C. Carr, 173-207. Kansas City: Westport Publishers.

Pilgram, T. 1987. Predicting Archaeological Sites from Environmental Variables, A Mathematical Model for the Sierra Nevada Foothills, California. In *BAR International Series 320*. Oxford: British Archaeological Reports.

Premo, L. S. 2001. A predictive model of Late Archaic Period site locations in the Tucson basin (Arizona). *MAI* 40 (01):162.

Scott, D. D. 1998. Euro-American Archaeology. In *Archaeology on the Great Plains*, ed. W. R. Wood, 481-510. Lawrence: University Press of Kansas.

Soil Survey Staff. 1981. Land Resource Regions and Major Land Resource Areas of the United States, 156. Washington, D.C.: United States Department of Agriculture, Soil Conservation Service.

Stancic, Z., and T. Veljanovski. 2000. Understanding Roman settlement patterns through multivariate statistics and predictive modelling. In *Beyond the Map: Archaeology and Spatial Technologies*, ed. G. Lock, 147-156. Amsterdam: IOS Press.

Tomlin, C. D. 1990. *Geographic Information Systems and Cartographic Modeling*. Englewood Cliffs: Prentice Hall.

Verhagen, P. 2000. Session 4 discussion: Archaeology, GIS and Cultural Resource Management. In *Beyond the Map: Archaeology and Spatial Technologies*, ed. G. Lock, 229-235. Amsterdam: IOS Press.

Warren, R. E. 1990a. Predictive Modelling in Archaeology: A Primer. In *Interpreting Space: GIS and Archaeology*, eds. K. M. S. Allen, S. W. Green and E. B. W. Zubrow, 90-111. London: Taylor & Francis.

———. 1990b. Predictive Modelling of Archaeological Site Location: A Case Study in the Midwest. In *Interpreting Space: GIS and Archaeology*, eds. K. M. S. Allen, S. W. Green and E. B. W. Zubrow, 201-215. London: Taylor & Francis.

Warren, R. E., and D. L. Asch. 2000. A Predictive Model of Archaeological Site Location in the Eastern Prairie Peninsula. In *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Wescott and R. J. Brandon, 5-25. London: Taylor & Fisher.

Wedel, W. R. 1963. The High Plains and Their Utilization by the Indian. *American Antiquity* 29 (1):1-16.

Wescott, K. L., and J. A. Kuiper. 2000. Using a GIS to Model Prehistoric Site Distributions in the Upper Chesapeake Bay. In *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*, eds. K. L. Wescott and R. J. Brandon, 59-72. London: Taylor & Francis.

Westcott, K. L., and R. J. Brandon, eds. 2000. *Practical Applications of GIS for Archaeologists: A Predictive Modeling Kit*. London: Taylor & Francis.

Wheatley, D. 1995. Cumulative Viewshed Analysis: A GIS-Based Method for Investigating Intervisibility, and its Archaeological Application. In *Archaeology and Geographical Information Systems: A European Perspective*, eds. G. Lock and Z. Stancic, 171-185. London: Taylor & Francis.

Wheatley, D., and M. Gillings. 2002. *Spatial Technology and Archaeology: The Archaeological Applications of GIS*. London: Taylor and Francis.

White, H. 2001. Discussion on southwestern Kansas archaeology.

Wood, W. R. 1998. Introduction. In *Archaeology on the Great Plains*, ed. W. R. Wood, 1-15. Lawrence: University Press of Kansas.

Appendix A

Logistic Regression Parameters for Southwest Kansas APM

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	20218	100.0
	Missing Cases	2	.0
	Total	20220	100.0
Unselected Cases		0	.0
Total		20220	100.0

a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Categorical Variables Codings

		Frequency	Parameter coding				
			(1)	(2)	(3)	(4)	(5)
sa_landforms	1	1363	1.000	.000	.000	.000	.000
	2	7991	.000	1.000	.000	.000	.000
	3	2192	.000	.000	1.000	.000	.000
	4	3886	.000	.000	.000	1.000	.000
	5	4501	.000	.000	.000	.000	1.000
	6	285	.000	.000	.000	.000	.000

Block 0: Beginning Block

Iteration History(a,b,c)

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	27070.131	-.434
	2	27069.895	-.441
	3	27069.895	-.441

a Constant is included in the model.
b Initial -2 Log Likelihood: 27069.895
c Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table[a,b]

			Predicted		
			binary		Percentage Correct
	Observed		0	1	
Step 0	binary	0	12301	0	100.0
		1	7917	0	.0
	Overall Percentage				60.8

[a] Constant is included in the model.

[b] The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.441	.014	935.376	1	.000	.644

Variable Names

Variable Name	SPSS Name
Distance to Intermittent Water	SA_D_INT
Distance to Perennial Water	SA_D_PER
Distance to Playa Lake	SA_D_PLA
Landforms (Categorical Variable)	SA_LANDF
	SA_LANDF(1)
	SA_LANDF(2)
	SA_LANDF(3)
	SA_LANDF(4)
	SA_LANDF(5)
Relief within 150m radius	SA_R150
Relief within 300m radius	SA_R300
Relief within 600m radius	SA_R600
Shelter Index within 300m radius	SA_SHR30
Slope	SA_SLOPE

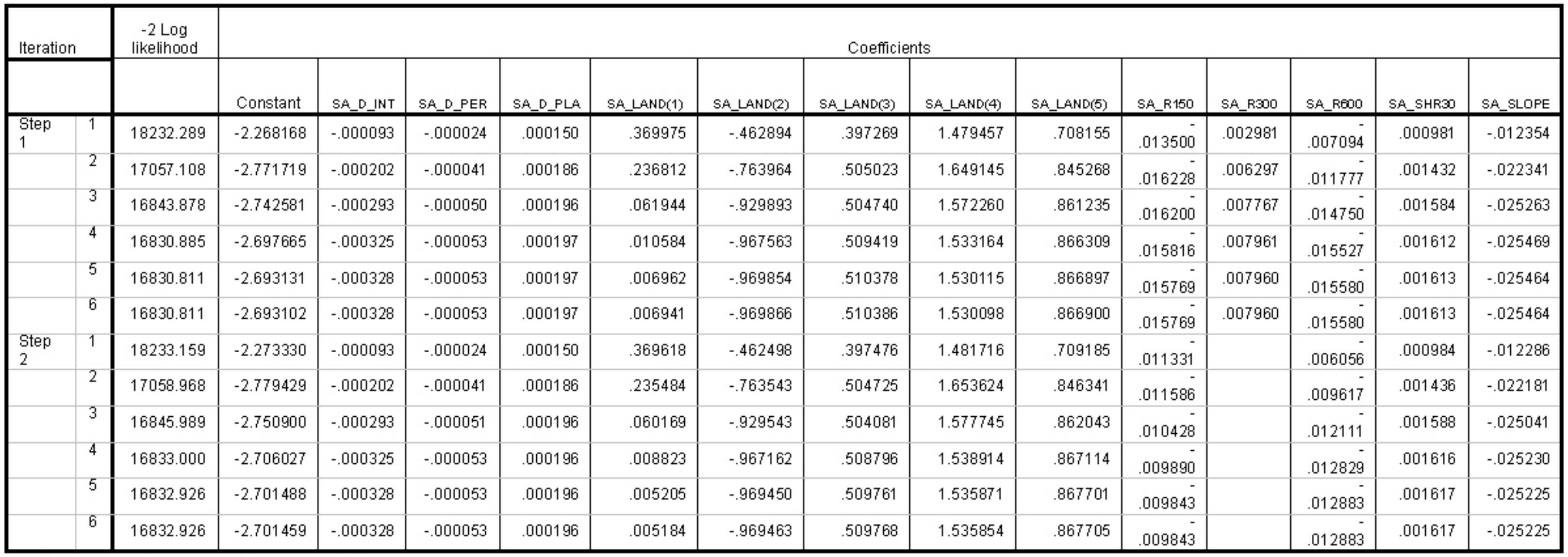
Variables not in the Equation(a)

			Score	df	Sig.
Step 0	Variables	SA_D_INT	2142.093	1	.000
		SA_D_PER	3571.891	1	.000
		SA_D_PLA	4969.536	1	.000
		SA_LANDF	5398.638	5	.000
		SA_LANDF(1)	38.711	1	.000
		SA_LANDF(2)	3646.964	1	.000
		SA_LANDF(3)	89.683	1	.000
		SA_LANDF(4)	3147.744	1	.000
		SA_LANDF(5)	570.284	1	.000
		SA_R150	1382.475	1	.000
		SA_R300	1704.509	1	.000
		SA_R600	1925.289	1	.000
		SA_SHR30	980.934	1	.000
		SA_SLOPE	949.502	1	.000

a Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Backward Stepwise (Conditional)

Iteration History(a,b,c,d)



Iteration		-2 Log likelihood	Coefficients													
			Constant	SA_D_INT	SA_D_PER	SA_D_PLA	SA_LAND(1)	SA_LAND(2)	SA_LAND(3)	SA_LAND(4)	SA_LAND(5)	SA_R150	SA_R300	SA_R600	SA_SHR30	SA_SLOPE
Step 1	1	18232.289	-2.268168	-.000093	-.000024	.000150	.369975	-.462894	.397269	1.479457	.708155	-.013500	.002981	-.007094	.000981	-.012354
	2	17057.108	-2.771719	-.000202	-.000041	.000186	.236812	-.763964	.505023	1.649145	.845268	-.016228	.006297	-.011777	.001432	-.022341
	3	16843.878	-2.742581	-.000293	-.000050	.000196	.061944	-.929893	.504740	1.572260	.861235	-.016200	.007767	-.014750	.001584	-.025263
	4	16830.885	-2.697665	-.000325	-.000053	.000197	.010584	-.967563	.509419	1.533164	.866309	-.015816	.007961	-.015527	.001612	-.025469
	5	16830.811	-2.693131	-.000328	-.000053	.000197	.006962	-.969854	.510378	1.530115	.866897	-.015769	.007960	-.015580	.001613	-.025464
	6	16830.811	-2.693102	-.000328	-.000053	.000197	.006941	-.969866	.510386	1.530098	.866900	-.015769	.007960	-.015580	.001613	-.025464
Step 2	1	18233.159	-2.273330	-.000093	-.000024	.000150	.369618	-.462498	.397476	1.481716	.709185	-.011331		-.006056	.000984	-.012286
	2	17058.968	-2.779429	-.000202	-.000041	.000186	.235484	-.763543	.504725	1.653624	.846341	-.011586		-.009617	.001436	-.022181
	3	16845.989	-2.750900	-.000293	-.000051	.000196	.060169	-.929543	.504081	1.577745	.862043	-.010428		-.012111	.001588	-.025041
	4	16833.000	-2.706027	-.000325	-.000053	.000196	.008823	-.967162	.508796	1.538914	.867114	-.009890		-.012829	.001616	-.025230
	5	16832.926	-2.701488	-.000328	-.000053	.000196	.005205	-.969450	.509761	1.535871	.867701	-.009843		-.012883	.001617	-.025225
	6	16832.926	-2.701459	-.000328	-.000053	.000196	.005184	-.969463	.509768	1.535854	.867705	-.009843		-.012883	.001617	-.025225

a Method: Backward Stepwise (Conditional)
b Constant is included in the model.
c Initial -2 Log Likelihood: 27069.895
d Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10239.084	13	.000
	Block	10239.084	13	.000
	Model	10239.084	13	.000
Step 2(a)	Step	-2.115	1	.146
	Block	10236.969	12	.000
	Model	10236.969	12	.000

a A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	16830.811(a)	.397	.539
2	16832.926(a)	.397	.538

a Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	825.883	8	.000
2	835.072	8	.000

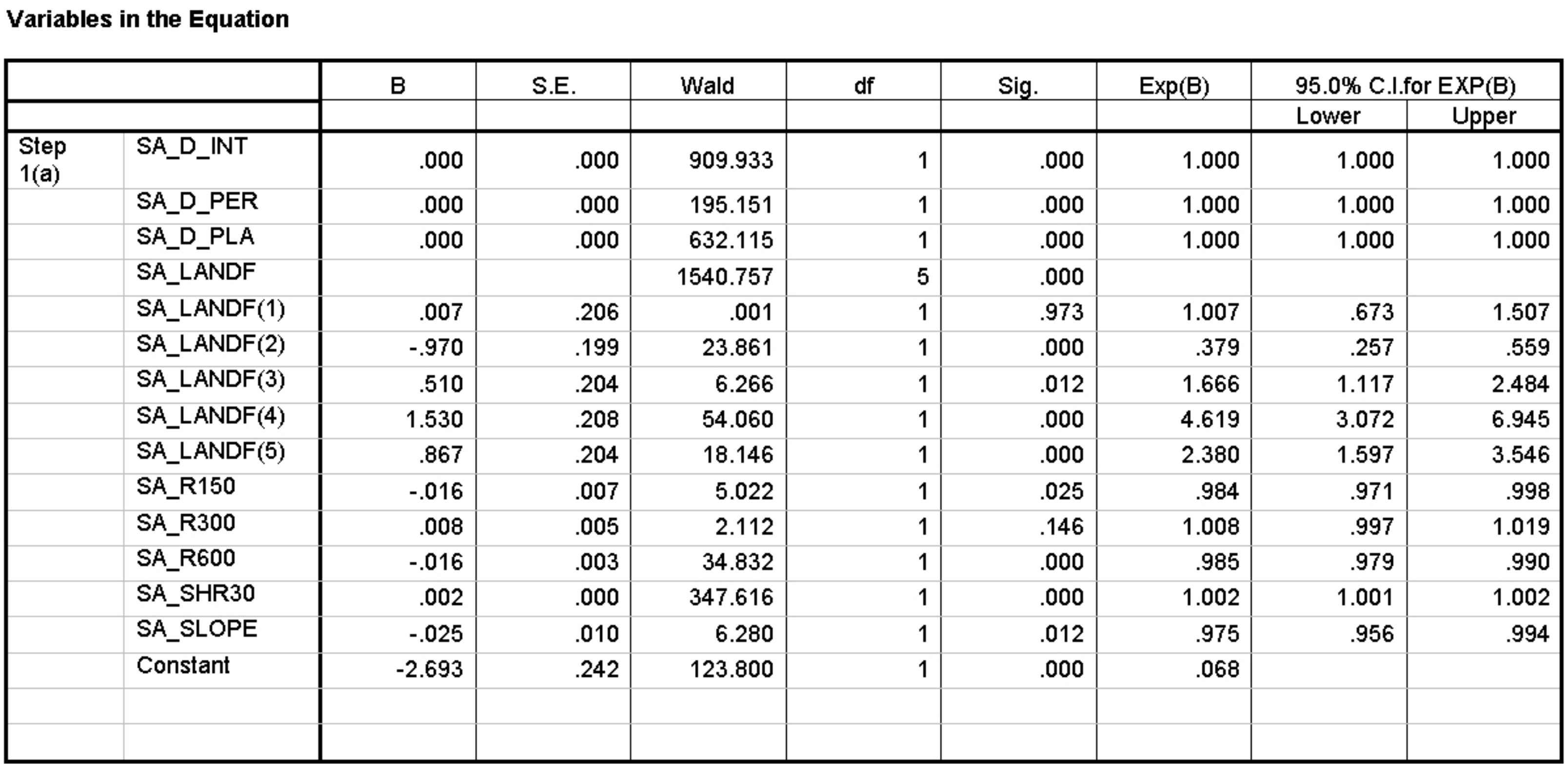
Contingency Table for Hosmer and Lemeshow Test

		binary = 0		binary = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2005	1999.440	17	22.560	2022
	2	1872	1924.538	150	97.462	2022
	3	1572	1845.282	450	176.718	2022
	4	1878	1731.121	144	290.879	2022
	5	1781	1553.851	241	468.149	2022
	6	1338	1268.621	684	753.379	2022
	7	823	904.861	1199	1117.139	2022
	8	649	546.881	1373	1475.119	2022
	9	228	344.557	1794	1677.443	2022
	10	155	181.848	1865	1838.152	2020
Step 2	1	2005	1999.416	17	22.584	2022
	2	1874	1924.530	148	97.470	2022
	3	1566	1845.287	456	176.713	2022
	4	1880	1730.937	142	291.063	2022
	5	1776	1553.440	246	468.560	2022
	6	1341	1268.574	681	753.426	2022
	7	832	905.782	1190	1116.218	2022
	8	645	546.477	1377	1475.523	2022
	9	235	344.535	1787	1677.465	2022
	10	147	182.022	1873	1837.978	2020

Classification Table[a]

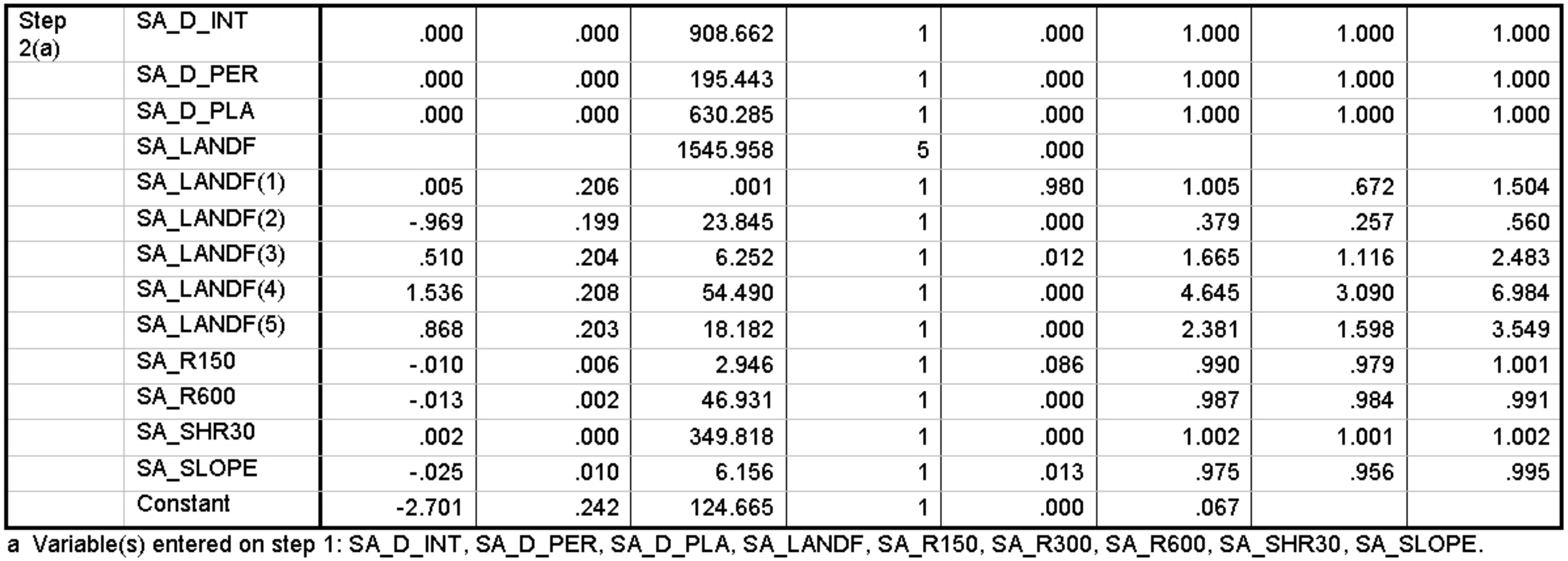
			Predicted		
			binary		Percentage Correct
	Observed		0	1	
Step 1	binary	0	10594	1707	86.1
		1	1876	6041	76.3
	Overall Percentage				82.3
Step 2	binary	0	10597	1704	86.1
		1	1875	6042	76.3
	Overall Percentage				82.3

[a] The cut value is .500



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	SA_D_INT	.000	.000	909.933	1	.000	1.000	1.000	1.000
	SA_D_PER	.000	.000	195.151	1	.000	1.000	1.000	1.000
	SA_D_PLA	.000	.000	632.115	1	.000	1.000	1.000	1.000
	SA_LANDF			1540.757	5	.000			
	SA_LANDF(1)	.007	.206	.001	1	.973	1.007	.673	1.507
	SA_LANDF(2)	-.970	.199	23.861	1	.000	.379	.257	.559
	SA_LANDF(3)	.510	.204	6.266	1	.012	1.666	1.117	2.484
	SA_LANDF(4)	1.530	.208	54.060	1	.000	4.619	3.072	6.945
	SA_LANDF(5)	.867	.204	18.146	1	.000	2.380	1.597	3.546
	SA_R150	-.016	.007	5.022	1	.025	.984	.971	.998
	SA_R300	.008	.005	2.112	1	.146	1.008	.997	1.019
	SA_R600	-.016	.003	34.832	1	.000	.985	.979	.990
	SA_SHR30	.002	.000	347.616	1	.000	1.002	1.001	1.002
	SA_SLOPE	-.025	.010	6.280	1	.012	.975	.956	.994
	Constant	-2.693	.242	123.800	1	.000	.068		



Step 2(a)	SA_D_INT	.000	.000	908.662	1	.000	1.000	1.000	1.000
	SA_D_PER	.000	.000	195.443	1	.000	1.000	1.000	1.000
	SA_D_PLA	.000	.000	630.285	1	.000	1.000	1.000	1.000
	SA_LANDF			1545.958	5	.000			
	SA_LANDF(1)	.005	.206	.001	1	.980	1.005	.672	1.504
	SA_LANDF(2)	-.969	.199	23.845	1	.000	.379	.257	.560
	SA_LANDF(3)	.510	.204	6.252	1	.012	1.665	1.116	2.483
	SA_LANDF(4)	1.536	.208	54.490	1	.000	4.645	3.090	6.984
	SA_LANDF(5)	.868	.203	18.182	1	.000	2.381	1.598	3.549
	SA_R150	-.010	.006	2.946	1	.086	.990	.979	1.001
	SA_R600	-.013	.002	46.931	1	.000	.987	.984	.991
	SA_SHR30	.002	.000	349.818	1	.000	1.002	1.001	1.002
	SA_SLOPE	-.025	.010	6.156	1	.013	.975	.956	.995
	Constant	-2.701	.242	124.665	1	.000	.067		

a Variable(s) entered on step 1: SA_D_INT, SA_D_PER, SA_D_PLA, SA_LANDF, SA_R150, SA_R300, SA_R600, SA_SHR30, SA_SLOPE.

Correlation Matrix

		Constant	SA_D_INT	SA_D_PER	SA_D_PLA	SA_LANDF(1)	SA_LANDF(2)	SA_LANDF(3)	SA_LANDF(4)	SA_LANDF(5)	SA_R150	SA_R300	SA_R600	SA_SHR30	SA_SLOPE
Step 1	Constant	1.000	-.076	-.171	-.054	-.774	-.765	-.746	-.738	-.752	.015	.022	-.040	-.566	-.016
	SA_D_INT	-.076	1.000	-.138	.083	.059	.051	-.039	.031	-.045	-.020	-.017	.117	-.057	.007
	SA_D_PER	-.171	-.138	1.000	.199	.098	-.008	.015	.053	.025	-.009	.003	.065	.034	-.004
	SA_D_PLA	-.054	.083	.199	1.000	-.107	-.113	-.139	-.136	-.218	-.014	.038	-.076	.027	-.014
	SA_LANDF(1)	-.774	.059	.098	-.107	1.000	.936	.911	.911	.927	-.009	.005	.004	.000	.003
	SA_LANDF(2)	-.765	.051	-.008	-.113	.936	1.000	.942	.939	.955	.008	-.002	-.019	-.027	.001
	SA_LANDF(3)	-.746	-.039	.015	-.139	.911	.942	1.000	.919	.941	.004	.002	-.027	-.011	-.005
	SA_LANDF(4)	-.738	.031	.053	-.136	.911	.939	.919	1.000	.944	-.007	-.018	-.064	-.012	-.010
	SA_LANDF(5)	-.752	-.045	.025	-.218	.927	.955	.941	.944	1.000	-.007	-.003	-.023	-.002	-.004
	SA_R150	.015	-.020	-.009	-.014	-.009	.008	.004	-.007	-.007	1.000	-.580	.085	-.012	-.558
	SA_R300	.022	-.017	.003	.038	.005	-.002	.002	-.018	-.003	-.580	1.000	-.702	-.029	-.016
	SA_R600	-.040	.117	.065	-.076	.004	-.019	-.027	-.064	-.023	.085	-.702	1.000	-.024	.042
	SA_SHR30	-.566	-.057	.034	.027	.000	-.027	-.011	-.012	-.002	-.012	-.029	-.024	1.000	.025
	SA_SLOPE	-.016	.007	-.004	-.014	.003	.001	-.005	-.010	-.004	-.558	-.016	.042	.025	1.000

Step 2	Constant	1.000	-.076	-.171	-.055	-.774	-.765	-.746	-.738	-.753	.034		-.035	-.566	-.016
	SA_D_INT	-.076	1.000	-.138	.085	.059	.051	-.039	.030	-.045	-.037		.148	-.058	.006
	SA_D_PER	-.171	-.138	1.000	.199	.098	-.008	.015	.053	.025	-.010		.095	.034	-.004
	SA_D_PLA	-.055	.085	.199	1.000	-.107	-.113	-.139	-.135	-.219	.010		-.069	.028	-.013
	SA_LANDF (1)	-.774	.059	.098	-.107	1.000	.936	.911	.912	.927	-.007		.011	.001	.003
	SA_LANDF (2)	-.765	.051	-.008	-.113	.936	1.000	.942	.939	.955	.008		-.029	-.027	.001
	SA_LANDF (3)	-.746	-.039	.015	-.139	.911	.942	1.000	.919	.941	.007		-.037	-.011	-.005
	SA_LANDF (4)	-.738	.030	.053	-.135	.912	.939	.919	1.000	.944	-.022		-.108	-.012	-.010
	SA_LANDF (5)	-.753	-.045	.025	-.219	.927	.955	.941	.944	1.000	-.011		-.035	-.002	-.004
	SA_R150	.034	-.037	-.010	.010	-.007	.008	.007	-.022	-.011	1.000		-.557	-.035	-.697
	SA_R600	-.035	.148	.095	-.069	.011	-.029	-.037	-.108	-.035	-.557		1.000	-.062	.044
	SA_SHR30	-.566	-.058	.034	.028	.001	-.027	-.011	-.012	-.002	-.035		-.062	1.000	.025
	SA_SLOPE	-.016	.006	-.004	-.013	.003	.001	-.005	-.010	-.004	-.697		.044	.025	1.000

Model if Term Removed(a)

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	SA_D_INT	-9149.435	1468.058	1	.000
	SA_D_PER	-8524.437	218.063	1	.000
	SA_D_PLA	-8746.025	661.239	1	.000
	SA_LANDF	-9279.567	1728.324	5	.000
	SA_R150	-8417.916	5.020	1	.025
	SA_R300	-8416.463	2.116	1	.146
	SA_R600	-8432.791	34.771	1	.000
	SA_SHR30	-8613.871	396.931	1	.000
	SA_SLOPE	-8418.545	6.280	1	.012
Step 2	SA_D_INT	-9149.650	1466.373	1	.000
	SA_D_PER	-8525.660	218.395	1	.000
	SA_D_PLA	-8746.130	659.334	1	.000
	SA_LANDF	-9284.206	1735.486	5	.000
	SA_R150	-8417.934	2.942	1	.086
	SA_R600	-8440.014	47.101	1	.000
	SA_SHR30	-8616.527	400.127	1	.000
	SA_SLOPE	-8419.541	6.156	1	.013

a Based on conditional parameter estimates

Variables not in the Equation

			Score	df	Sig.
Step 2(a)	Variables	SA_R300	2.112	1	.146
	Overall Statistics		2.112	1	.146

a Variable(s) removed on step 2: SA_R300.