CHAPTER TWO

PREDICTIVE PROBABILITY MODELS

Conceptual Framework

Proper management of cultural resources should not only entail protection and interpretation, but also include research into understanding and explanation of historic phenomenon. Archaeologists are interested in objects of the past, and also why those objects are found in certain locations and not others. This knowledge requires understanding interactions of past peoples with their environments.

The location of archaeological sites exhibit non-random tendencies or patterning throughout a landscape (Parker 1985). This patterning is a result of past people's tendency to interact with the landscape in "favorable" settings. Favorable settings refer to sites that are preferred over other locations because of specific landscape characteristics (e.g., proximity to navigable water, access to trade routes, prominent setting with high visibility). Certain variables, either environmental or social, within the landscape can produce patterning. It is these non-random redundancies that predictive models exploit in attempt to explain the relationship between locational characteristics and archaeological sites. Predictive models make use of existing knowledge to anticipate or predict events that are to come or that have already taken place.

An archaeological predictive model is essentially a map that indicates the relative potential of encountering an archaeological site. These maps are often referred to as "sensitivity" maps because they indicate the sensitivity of one location in relation to another for the presence of cultural resources (Parker and Johnson 1986). Predictive modeling emerged over the last few years as an important component of archaeological research (Carr 1985, Kohler and Parker 1986). "Most archaeological predictive models rest on two fundamental assumptions. First, the settlement choices made by the prehistoric [historic] peoples were strongly influenced or conditioned by characteristics of the natural environment. Second, the environmental [or social] factors that directly influence these choices are portrayed, at least indirectly, in modern maps of environmental [or social] variation across an area of interest" (Allen, Green, and Zubrow 1990b, 62). With these assumptions fulfilled, predictive models hold tremendous potential as planning tools. In the long run, they can reduce costs for archaeological survey, mitigation, and clearance (Minnesota Department of Transportation 1996). For example, the Minnesota Department of Transportation (1996) prepared an archaeological predictive model in order to avoid areas of high sensitivity in future construction projects. Warren (1987) also constructed a predictive model, but with the goal of conserving time and resources in locating archaeological sites in the Western Shawnee National Forest.

Warren noted that one of the most powerful approaches to prediction is a family of procedures called probability models (Allen, Green, and Zubrow 1990a). Probability models work under the assumption that there is data on positive and negative responses to stimuli (dichotomy). In other words, the dependent variable is either a positive or negative (success vs. failure, presence vs. absence) response with respect to the independent variable(s). Like linear regression, the independent variable(s) are predictors of the dependent variable. "Probability-based predictions have several advantages over predictions on other scales. For instance, they are readily interpretable (values range between 0 and 1), they can be treated as ratios (a probability of 0.6 is twice as high as a probability of 0.3), and their accuracy can be tested with sample data" (Allen, Green, and Zubrow 1990a, 93).

Logistic Regression Analysis

There are many statistical approaches to predictive probability modeling. However, the most popular of these is the logit or logistic regression model (Allen, Green, and Zubrow 1990a). In Carr's (1985) "Introductory Remarks on Regional Analysis", five popular techniques (i.e., density transfer, density regression, significance regression, discriminant function analysis, and logistic regression analysis) of predictive modeling were critically scrutinized. Logistic regression, although very similar to discriminant function analysis, was less constrained by statistical assumptions[1]. It was also found to provide more powerful and consistent predictions when the aforementioned statistical assumptions were violated (Kvamme, 1983; Press and Wilson 1978). In addition, logistic regression analysis readily accepts mixtures of nominal, ordinal, interval, and ratio scaled independent variables. Use of logistic regression was also scrutinized and wholeheartedly supported by Parker (1985) in her article: "Predictive Modeling of Site Settlement Systems Using Multivariate Logistics." If one considers all these advantages, plus the fact that the resulting formula from logistic regression is easily interpreted, logistic regression becomes the clear choice for use in archaeological prediction models.

Logistic regression employs the use of independent variables to create a mathematical formula that predicts the probability that a site occurs on any give parcel of land (Allen, Green, and Zubrow 1990b). The key to logistic regression is that the dependent variable is dichotomous. Unlike multiple regression, which predicts scores for a continuous dependent variable, logistic regression predicts the probability of membership in one of the available groups (i.e. site/non-site). The independent variable(s) in this model are predictors of the dependent variable and can be measured on nominal, ordinal, interval, or ratio scale. The relationship between the dependent variable

[1] Discriminate function analysis:
(1) Assumes multivariate normality of data
(2) Assumes equal covariance matrices
(3) Does not readily accept mixtures of categorical and interval-scale independent variables

and the independent variable(s) is nonlinear. It is this relationship that is utilized to predict the probability of group membership for each case in the model.

The standard logistic regression formula for a model with multiple independent variables is:

$$p(\mathrm{B}) = \frac{Exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i)}{1 + Exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i)}$$

Or simplified

$$p(\mathrm{B}) = \frac{1}{1 + Exp(-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i))} \qquad \textbf{2.1}$$

Where $p(\mathrm{B})$ is the probability (p) that case "i" is a member of group B, such that $p(\mathrm{B}) = 1$ (i.e. site presence); *Exp* is a function that raises the number *e* exponentially to the power of the value enclosed in parentheses, where the number *e*, Euler's number, is the irrational number whose natural logarithm is 1 ($\ln(1) = 2.71828\ldots$); α is the intercept constant; the βs are the coefficients for the independent variables; and the "x"s are the independent variables for the corresponding β coefficient.

The α parameter, called the intercept, represents the value of the dependent variable (Y) when x is zero. The parameter β represents the change in Y associated with one-unit increase in x, or the slope of the line that provides the best linear estimate of Y from x. In linear regression, the least squares method is most often use to estimate parameters (i.e., α and β). This method selects those values of α and β which minimize the sum of squared deviations of the observed values of Y from the predicted values based upon the model. To estimate α and β coefficients for the independent variables in the logistic regression model, two methods are commonly used: the maximum-likelihood and the least-squares regression fitting procedures (Warren 1987). Unlike linear regression, the least-squares regression approach is plagued with many statistical problems, so the maximum-likelihood fitting procedure is most frequently used (Hosmer and Lemeshow 1989). Although the maximum-likelihood method requires a complex series of iterations in which trial coefficients are proposed, tested, and refined to find an optimum solution, current statistics software and computer hardware make this ideal approach feasible. In general, the maximum-likelihood technique is used to maximize the log-likelihood function, which indicates how likely it is to obtain the observed values of Y, given the values of the independent variables and parameters (i.e., α and β) (Menard 1995).

Probabilities produced from the logistic procedure are used to derive the dichotomous dependent variable for each location. To accomplish this, a *cutpoint* (*c*) value must be selected to delineate sites from non-sites. Each location's probability is compared to this value (*c*) to determine membership. If the estimated probability exceeds or is equal to the cutpoint value then the location is considered to be a site (if $p(B) >= c$, then = 1), otherwise it is considered to be a non-site (if $p(B) < c$, then = 0).

Stepwise Logistic Regression

The term "stepwise" refers to the use of decisions made by computer, rather than choices made by the researcher, to select a set of predictors (i.e., independent variables) for inclusion in or removal from the logistic model. Stepwise logistic regression is most often used in situations were the "important" independent variables are not known and associations with the outcome not well understood (Hosmer and Lemeshow 1989). In these instances, most studies will collect many possible independent variables and screen them for significance. Stepwise logistic regression offers a fast and effective means of screening a large number of variables, and simultaneously fit a number of logistic regression equations.

Opponents of the use of stepwise regression criticize the technique as an admission of ignorance on the phenomenon being studied (Studenmund and Cassidy 1987). There is also general agreement that the use of stepwise techniques is inappropriate for theory-testing because it capitalizes on random variations in the data and produces results that tend to be idiosyncratic and difficult to replicate in any sample other that the sample in which they were originally obtained (Menard 1995). However, stepwise proponents support its use in purely predictive and exploratory research (Menard 1995). Both Hosmer and Lemeshow (1989) and Agresti and Finlay (1986) support the stepwise logistic regression technique as a useful and effective data analysis tool in these situations. To mention a few, stepwise logistic regression was utilized to uncover settlement patterns of ancient civilizations in the country of Jordon (Christopherson, Guertin, and Borstad 1996), model prehistoric site locations near Pinyon Canyon, Colorado (Kvamme 1984), and is also currently being used to identify archaeological sites in the state of Minnesota (Minnesota Department of Transportation 1996).

There are two basic forms of stepwise logistic regression: forward inclusion and backward elimination. In forward logistic regression all independent variables are initially withheld from the model. At subsequent steps in the procedure, those variables determined to be significant are added to the model while all others are withheld. Just the opposite occurs in backward logistic regression in which all independent variables are initially include in the model. At subsequent steps in the procedure, those variables determined insignificant are eliminated from the model until the remaining variables are all deemed "important."

In stepwise logistic regression (e.g., forward or backward), selection or deletion of variables from the model is based on a statistical algorithm that checks for "importance" of variables, and either includes or excludes them on the basis of a fixed decision rule. The likelihood ratio chi-square test is used to assess significance in logistic regression since the errors are assumed to follow a binomial distribution. This test assigns a *p*-value to each variable to assess significance. Therefore, the most important variable is the one with the smallest *p*-value.

An important element of stepwise logistic regression is selection of removal and entry criteria (e.g., fixed decision rule) to determine variable significance. The removal criterion (*p*R) is the *p*-value (i.e. probability value) used to eliminate insignificant independent variables. If a variables *p*-value is equal to or greater than this number it will be eliminated from the model. The entry criterion (p_E) value determines which independent variables will be included in the model. If a variable's *p*-value is less than this value then it will be entered into the model.

The following subsection is a simplified description of forward stepwise logistic regression. This example is included to clarify the stepwise process for the reader. It should be noted, as mentioned earlier, that the processes (i.e., steps) of backward stepwise logistic regression are essentially the same. However, in backward elimination all independent variables are included in the initial model, and are then evaluated for "importance" at subsequent steps.

Forward Stepwise Logistic Regression

Dependent Variable:	Absence or Presence Union Civil War Forts
Independent Variables:	Elevation, Slope, and Distance from Confederate Forts
p_E:	0.20
*p*R:	0.15

STEP 0

1. Fit the "intercept only" logistic regression model
    - Compute the log-likelihood for this model
2. Fit each independent variable (i.e., "Elevation", "Slope", and "Distance from Confederate Forts") into this model separately
    - Compute log-likelihood values for each independent variable
    - Perform the likelihood ratio tests on each variable
    - Compute *p*-values for each independent variable (i.e., 0.0037, 0.6521, and .0071)
3. Select the independent variable with the smallest *p*-value: "Elevation" (i.e., 0.0037)
4. Proceed to Step 1 if 0.0037 (*p*-value of "Elevation") < 0.20 (*p*E) otherwise STOP

Step 1

1. Fit the "Elevation" plus the intercept logistic regression model
    - compute the log-likelihood for this model
2. Fit each remaining independent variable into this model (i.e., "Slope" and "Distance from Confederate Forts")
    - Compute log-likelihood values for each independent variable
    - Perform the likelihood ratio tests on each variable
    - Compute *p*-values for each independent variable (i.e., 0.6631 and 0.0070)
3. Select the independent variable with the smallest p-value: "Distance from Confederate Forts" (i.e., 0.0070)
4. Proceed to Step 2 if 0.0070 (*p*-value of "Distance from Confederate Forts") < 0.20 (p_E) otherwise STOP

Step 2

1. Fit the "Elevation, Distance from Confederate Forts", plus the intercept logistic regression model
    - Compute the log-likelihood for this model
2. Check for removal/elimination of independent variables in the model (i.e., "Elevation" and "Distance from Confederate Forts")
    - Compute log-likelihood without "Elevation" and "Distance from Confederate Forts"
    - Perform the likelihood ratio test on "Elevation" and "Distance from Confederate Forts"
    - compute *p*-values for "Elevation" and "Distance from Confederate Forts" (i.e., 0.0000 and 0.0003)
3. Select the independent variable ("Distance from Confederate Forts"), that when removed, yields the largest *p*-value (i.e., 0.0003)
4. If 0.0003 (*p*-value of "Distance from Confederate Forts") > 0.15 (p_R) then remove "Distance from Confederate Forts", else "Distance from Confederate Forts" remains in the model
5. Fit each remaining independent variables into this model (i.e., "Slope")
    - Compute log-likelihood values for each independent variable
    - Perform likelihood ratio tests for each independent variable
    - Compute *p*-values for each independent variable (i.e., 0.3344)
6. Select the independent variable ("Slope") with the smallest p-value (i.e., 0.3344)
7. Proceed to Step 3 if 0.3344 (*p*-value of "Slope") < 0.15 (p_E), otherwise STOP

Step 3

This procedure is identical to Step 2. The program checks for backward elimination followed by forward selection. This process continues until the last step, Stop Step.

Stop Step

This step occurs when:
1. all variables have been entered into the model, or
2. all variables in the model have *p*-values < p_R (0.15) and the variables not included in the model have *p*-values > p_E (0.20)

All variables in the model at this final step are important relative to the criteria of p_R and p_E.

In some instances an independent variable may illustrate the *suppressor* effect, that is, when it appears to be statistically significant only when another variable is controlled or held constant (Agresti and Finlay 1986). The disadvantage of using forward logistic regression is the possible exclusion of variables involved in the suppressor effect (Menard 1995). Although Menard (1995) mentioned that both forward and backward stepwise techniques often generate identical results; backward elimination is more likely to uncover these relationships since all variables are initially included in the model.

Geographic Information Systems and Predictive Probability Models

> Archaeology and geography are sister disciplines. They are both concerned with the patterning of human activities in space and time
>
> P. Peregrine, 1988. *Geographic Information Systems in Archaeological Research: Prospects and Problems*, 874

Archaeologists have extensive spatial data handling requirements (Reilly and Rahtz 1992). Archaeology deals with the unique position in space and time of phenomenon and the latent relationships existing between them. Since early this century archaeologists have used a variety of techniques to visualize, analyze, and interpret spatial patterning, but maps have always been an ideal media (Reilly and Rahtz 1992). Therefore, it is no surprise that archaeologists have paid close attention to developments is spatial analysis within geography, most notably Geographic Information Systems (GIS).

A GIS is a collection of computer hardware and software, spatial data, and personal that efficiently capture, store, manipulate, analyze create, and display geographically referenced data. GIS incorporates computer cartography and a relational database into one package, that is, every mapped feature is linked to a record in a tabular database which holds its attributes.

GIS data sets are organized into map layers, in which each layer contains different information on the same geographic area. These same map layers share a common coordinate system that allows them to share the same geographic space. In the same way, each database associated with these map layers has a *key field* that links each database together. It is this linkage between the mapped feature and the database that makes analysis of geographic data possible (Figure 2.1). With such "intelligent" maps, spatial queries, measurements, and more complex problems can be answered easily.

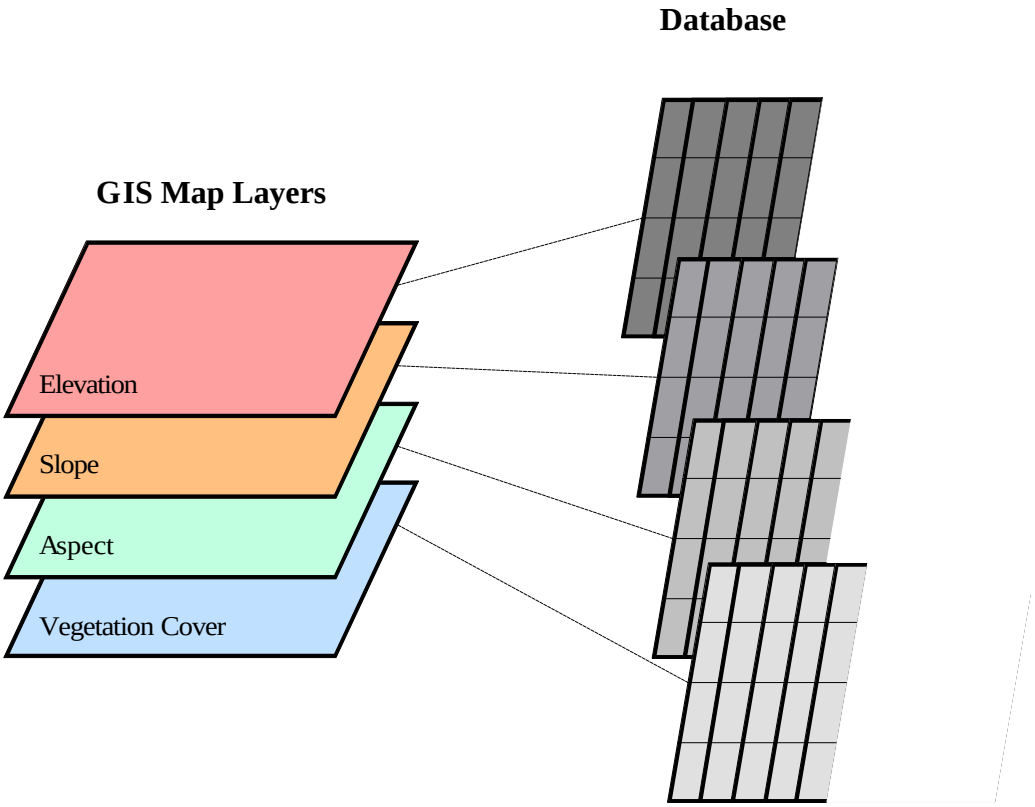

Figure 2.1. Generalization of GIS data structure.

Predictive models in archaeology fall into the general category of cartographic modeling. Traditionally these models were developed using analog maps and manual map overlay procedures. The advent of GIS has greatly simplified this process. Spatial data from large regions can be digitized and geo-referenced within a GIS to allow rapid investigation of relationships between the locations of sites and the environment. Kvamme stated "… development and application of models of regional archaeological distributions is greatly facilitated through GIS (Allen, Green, and Zubrow 1990c, 112)."