COENOSES 12(1):1-21, 1997
© C.E.T.A., Goriza, Italy

RESOURCE SELECTION BY ANIMALS: THE STATISTICAL ANALYSIS OF BINARY RESPONSE[1]

Daniel Campos[1], Amarjot Kaur[1], G. P. Patil[1], W. J. Ripple[2], and C. Taillie[1]

1 Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802 USA
2 Environmental Resources Sensing Applications Laboratory (ERSAL), Department of Forest Resources, Oregon State University, Corvallis, OR 97331 USA

Keywords: Data (binary, observational); Deviance; Dispersion (over, under); Distribution (Bernoulli, binomial, normal); Fernbirds; Generalized Linear models; Habitat (fragmentation, selection); Logistic regression; Maya settlements; Northern spotted owl; Quasi-likelihood; Resource (availability, use); Resource selection (probability functions, studies); Retrospective studies; Sampling (designs, probabilities, protocols).

Abstract: In order to survive, animals must select among the resources available in their environment. Resource selection studies aim to identify the habitats and prey items that different animal species select for use among available choices. These studies have progressively become more reliant on statistical methods for analysis. As in other ecological applications, classical linear models are somewhat constraining due to the assumptions of constant error variance and normal distribution for the response variable Y. Generalized linear models (GLM) add necessary flexibility to model ecological data by allowing Y to come from an exponential family and by taking the variance of Y to be proportional to a variance function $V(\bullet)$ of the mean. The variance function $V(\bullet)$ determines whether there is an exponential family compatible with the moment model. If there is one, maximum likelihood can be used for inference. If there is not one, quasi-likelihood is an alternative method for inference that does not make distributional assumptions for the response variable.

In this paper, an introduction to resource selection studies is presented; it includes definitions, motivation, sampling designs, and analysis techniques based on resource selection probability functions. GLM and quasi-likelihood are also introduced and applied in the analysis of three different resource selection studies with binary response variables.

1. Introduction

Resource selection studies attempt to identify the habitats and food items that different animal species select for use among available choices. Statistical science has become an intrinsic part of such studies. Researchers in biology, ecology, and related fields rely on statistical analysis to draw conclusions about resource selection by animals. These studies, however, have not always relied on statistics. Early researchers simply described the use and availability of resources in their studies. Resource selection was first quantified in the analysis of food studies, and Scott (1920) is cited as the first author to quantify selection by developing a ratio of the rate of prey consumption by fish to the density of the prey's presence (Manly et al. 1993, p. 9; Cock 1978). Since the early studies, resource selection research has become progressively more in tune with modern statistical analysis as the latter has been developed. Manly et al. (1993) comprised a unified statistical theory for analysis of resource selection studies based on the concept of a resource selection probability function.

The class of generalized linear models (GLM), an extension of classical linear models, is one of the most important contemporary tools for analysis of resource selection. In ecological studies, classical linear models (linear regression and analysis of variance) are somewhat constraining due to the assumptions of constant error variance and normal distribution for the response variable. Breakdown of these assumptions is traditionally corrected by transforming the response variable. However, a transformation $g(Y)$ of the response has some limitations. First, $g(y)$ may not be defined for all values y of the response in the data set; for example, log(0), logit(0), and logit(1) are not defined. Second, the variance of Y may depend upon a design parameter m; for instance, the variance of a sample proportion is expected to be inversely proportional to the sample size m, but after transformation, the dependence of the variance of $g(Y)$ upon

1 Prepared with partial support from the Statistical Analysis and Computing Branch, Environmental Statistics and Information Division, Office of Policy, Planning, and Evaluation, United States Environmental Protection Agency, Washington, DC under a Cooperative Agreement Number CR-821531. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred. We are thankful to Bryan Manly and László Orlóci for their comments and suggestions on the earlier draft of the paper.

m can only be expressed approximately and often depends on unknown parameters. Third, linearizing transformations may fail to stabilize the variance (Kaur et al. 1995).

Generalized linear models add necessary flexibility for modeling ecological data by relaxing classical assumptions. This added flexibility can be illustrated as follows. Let $\mathbf{x} = (x_1,...,x_p)$ be a column vector of covariates and let the response $\mathbf{Y} = (Y_1,...,Y_n)$ be a vector of independent, normally distributed random variables with constant variance σ^2 and mean $\mu=(\mu_1,...,\mu_n)$ so that

$$E(Y_i) = \mu_i = \beta' x = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i=1,...,n,$$

where $\beta'=(\beta_1,...,\beta_p)$ is a row vector of unknown parameters. This model can be broken into three parts:

(a) Random component: $\mathbf{Y}$ is independent, normally distributed with constant variance σ^2 and

$$E(\mathbf{Y}) = \mu. \tag{1}$$

(b) Systematic component: the covariates such that

$$\eta = \sum_{j=1}^{p} \beta_j x_j.$$

(c) Link function between random and systematic components in linear models:

$$g(\mu) = \mu = \eta.$$

Classical linear models require a normal distribution for $\mathbf{Y}$ in part (a) and an identity link function in part (c). Generalized linear models extend the classical model by allowing the response $\mathbf{Y}$ in part (a) to come from an exponential family and by letting the link function in part (c) be any monotonic differentiable function $g(\bullet)$ (McCullagh & Nelder 1983, p. 19-20). The link function for a generalized linear model takes the form

$$g(\mu) \equiv \eta = x_1\beta_1 + x_2\beta_2 + ... + x_p\beta_p. \tag{2}$$

The GLM framework, therefore, applies a transformation $g(\mu)$ to the mean. Furthermore, the variance of $\mathbf{Y}$ is not taken to be the constant σ^2; instead, it is proportional to a variance function $V(\bullet)$ of the mean μ so that

$$\text{var}(\mathbf{Y})=\varphi V(\mu). \tag{3}$$

The basic moment model for the class of generalized linear models is given by Equations 1, 2 and 3. The form of the variance function $V(\bullet)$ determines whether there is a linear exponential family that is compatible with the moment model. If such a family exists and the data come from that family, maximum likelihood can be used for inference; however, the necessary linear exponential family does not always exist, especially if over-dispersion (or under-dispersion) is present in the data. In such cases, quasi-likelihood is a useful method for inference that does not make distributional assumptions for the response variable. If the moment model corresponds to a linear exponential family, quasi-likelihood estimates are the same as maximum likelihood estimates.

Due to the characteristics explained above, generalized linear models and quasi-likelihood methods comprise an important framework for analysis in ecology, and their importance is extended to resource selection studies. This paper is concerned with the theory and applications of GLM and quasi-likelihood to binary response resource selection studies. These are common studies where the response variable is the use or rejection of a resource unit.

Here we will examine three binary response studies. The first two examples concern nest selection by birds. The northern spotted owl *(Strix occidentalis caurina)* is an endangered species threatened by exploitation of its old forest habitat in Oregon, USA. We review and expand a study that attempts to determine how much forest fragmentation the owl can tolerate for nesting. A similar study of fernbirds in Otago, New Zealand, aims to identify significant variables that influence nest selection. The third study is an analysis of the selection of settlement sites in Belize by prehistoric Maya communities. This is a relevant example since, in the study of the use of resources by human societies, the researcher is confronted with the same basic question of animal selection studies: which resources are selected for use among available choices?

Section 2 provides a general introduction to resource selection studies that includes definitions, types of data, sampling designs, motivations, and the important concept of a resource selection probability function. Section 3 introduces logistic regression as a technique for analysis of binary data within the GLM framework. Section 4 presents quasi-likelihood methods and their applications to sample studies, especially in the case of over-dispersed binomial data. The final Section 5 includes a general discussion of the issues and examples presented in preceding sections.

2. Resource Selection Studies

In this section, we introduce definitions, motivations, nature of the data, sampling designs, sampling protocols, and probability functions for resource selection studies. We follow closely the discussion that Manly et al. (1993) bring forth in their introduction to selection studies.

2.1 Definitions

Adequate quantities of usable resources are essential to sustain animal populations. Therefore, biologists are often interested in identifying resources used by animals and recording the availability of those resources. Studies that attempt to determine which resources animals select among available choices are called resource selection studies. They provide important information about the nature of animals and their needs for survival (Manly et al. 1993, p. 1).

The use of a resource is the quantity of the resource used by an animal (or population of animals) in a fixed period of time. Resource availability is defined as the quantity accessible to an animal (or population of animals) during the same period of time. Selection is the process by which an animal chooses to use a resource. When resources are used dis-

proportionately to their availability, use is said to be selective (Manly et al. 1993, p. 1; Thomas & Taylor 1990).

Resource selection occurs at different hierarchical levels: the geographic range of a species, the individual home range within a geographical range, general features or habitats within the home range, and food items within general features or feeding sites. In order to make inference, the researcher must consider the level at which selection is studied since selection criteria may be different at each level (Manly et al. 1993, p. 1; Johnson 1980).

We assume that a species selects resources that best satisfy its requirements, and that high quality resources are preferred to low quality ones. The availability of resources, however, is not generally uniform. Therefore, in order to reach valid conclusions about resource selection, used resources must be compared to available or unused resources (Manly et al. 1993, p. 1).

The most common selection studies are interested in food or habitat selection. Food selection may occur among different species of prey or among different sizes, colors, and other characteristics of the same prey species. Habitat selection may occur among several discrete habitat categories such as forests or open fields or among continuous variables such as vegetation density and distance to water. Selection studies, then, may include discrete or continuous variables or a combination of both (Manly et al. 1993, p. 2).

2.2 Motivation

There are several different situations in which resource selection studies are relevant. In some cases, it is necessary to assess the long term resource requirements for survival of an animal species. For example, there is a heated debate in the Pacific Northwest about the exploitation of old forest forests and its impact on the survival of the northern spotted owl. Ripple et al. (1991) and Ramsey et al. (1994) have conducted studies in an attempt to determine how much old forest is necessary for the survival of the species. This particular example is discussed in detail in subsequent sections. Resource selection studies may also be conducted, among other purposes, to model and project the impact of habitat change and to evaluate the effect of domestic animals on wild animal forage (Manly et al. 1993, p. 3).

2.3 The Data for Resource Selection Studies

The data used in resource selection studies may be gathered through a census or through one or more sampling methods. The data may also be categorical or continuous, and may be univariate or multivariate. It is assumed that the resource under study consists of a number of discrete units, and the set of all units is called the universe of available resource units (Manly et al. 1993, p. 4). The division of the resource into units may occur naturally; for example, when resource units are individual prey items or tree cavities for nesting for an owl species. In other cases, the division into units is imposed by the researcher. For example, the researcher divides the study area into grid sections or plots as in Ryder's (1983) study of winter habitat selection by pronghorn in Wyoming. Ryder divided the study area into plots of land where the presence or absence of antelope was recorded during winter.

Studies may classify observed resource units into categories or specific variables may be measured for resource units (Thomas & Taylor, 1990). For example, Murphy et al. (1985) collected habitat association data for white-tailed deer by classifying observations of deer in six different habitat types, while Ramsey et al. (1994) studied habitat selection of the flammulated owl by measuring a combination of eighteen continuous variables and two categorical factors on each nesting site (or resource unit).

Furthermore, individual animals may be identified in selection studies. Gionfriddo & Krausmann (1986) identified individual radio-collared sheep to study their summer habitat selection. Sometimes, however, data collection methods may not distinguish among individuals as in Keating's (1985) use of bighorn sheep pellets to study their winter food habits (Thomas & Taylor 1990).

2.4 Sampling Designs

According to Johnson (1980), once the researcher has decided to study resource selection for a particular animal species, he must decide on the scale of selection he desires to study. It is essential to consider the knowledge on the biology of the animal. For example, the scale of the study is different for territorial and non-territorial animals, and this has a significant impact on which resources may be assumed to be available to the animal at the scale of interest (Manly et al. 1993, p. 5).

In designing a selection study, the researcher must also decide on the choice of study area and its boundaries. To choose a study area, the researcher should consider the distribution of resource units, the scale of selection studied, the resources that are truly available to the animals, and the budget and technical constraints for sampling.

Thomas & Taylor (1990) identify and describe three general field study designs to investigate resource selection. These designs differ in whether resource use and availability are measured at the population level or for each animal.

Design I

Measurements occur at the population level. Used, unused, and available resources are estimated for the collection of all animals in the whole study area; no individual animals are identified. For example, when Stinnett & Klebenow (1986) examined cover-type selection of California quail, flushes observed during ground surveys were classified into cover types; maps and aerial photographs were partitioned into the respective cover types to evaluate availability. Individual animals were not identified; use and availability were measured at the population level.

Design II

Individual animals are collected or identified with neck collars, tags, or radio transmitters. Use of resources is estimated for each animal, but available resources are measured for the population of all animals. Design II studies often compare

the relative number of locations of marked individuals in each habitat type to the proportion of the respective habitat types in the study area. For example, Roy & Dorrance (1985) compared habitat use by individual coyotes in their home ranges to habitat availability in the entire study area.

Design III

Individual animals are identified and their use of resources is estimated; in this design, however, available resources are measured for each animal and not at the population level. These studies often identify home ranges for individual animals and compare use and availability of resources within that range. Rolley & Warde (1985), for instance, used radio telemetry to identify home ranges (territories) for individual bobcats and measured individual use of various vegetation types.

Designs II and III involve uniquely identified individuals. Thus, in order to make inferences about the animal species population, the researcher must assume that the identified animals are a random sample from the population. This assumption requires a sample design with multiple stages: selecting a sample of individual animals, selecting a sample of used and available resource units for each animal, and sometimes, sub-sampling from the chosen resource units (Manly et al. 1993, p. 7).

Furthermore, designs II and III allow the analysis of resource selection for individual animals. The estimates calculated from observations on individuals may be used to estimate population parameters and to obtain estimates of variability for parameter estimates. Manly et al. (1993) suggest the following advantages in this approach:

> (a) The observations on an individual animal may be time-dependent. For instance, in habitat selection studies, relocations of one animal may depend on the time of the day, and in food studies, selection of consecutive prey items may be dependent. Therefore, it is better to estimate variances and test hypothesis based on variation between animals rather than variation between observations on one animal. In this way, inference relies on random sampling (design-based inference) rather than on the assumption that the statistical model is correct (model-based inference).
>
> (b) With designs II and III, estimation techniques that are applicable at the population level for design I become possible for individuals.
>
> (c) Variation among individuals can be examined to determine different selection behaviors among animals such as age or sex differences and unusual selectivity by an individual.

Note that the three field study designs previously discussed are for observational studies since controlled experimental studies are difficult to carry out in a wildlife setting; nevertheless, a few experiments also have been conducted to determine resource selectivity (Manly et al. 1993, p. 8).

2.5 Sampling Protocols

Resource selection may be assessed by comparing any two of three possible sets of resource units: used, unused or available units. Each combination of sets measured yields a different sampling protocol (Manly et al. 1993, p. 8). Sampling protocol A consists of a census or a random sample of available resource units and a random sample of used resource units. Under sampling protocol B, a census or a random sample of available units is also obtained, but in this case, a random sample of unused resource units is taken. With sampling protocol C, independent samples of used and unused resource units are taken.

Each of these sampling protocols may be utilized for any of the three sampling designs described in the previous section. Each particular combination of design and protocol used to collect data has different implications on the underlying assumptions for subsequent data analysis, for example, on whether the availability of resources is the same for all animals (Manly et al. 1993, p. 9).

2.6 Resource Selection Probability Function

Manly et al. (1993) attempt to provide an unified statistical theory for resource selection studies based on the concept of a resource selection function. This function depends on several variables measured on a resource unit and takes a value proportional to the probability of that unit being used. This analytical approach may be applied whenever the resource consists of an universe of N available (used or unused) units, and where several characteristics $\mathbf{x} = (x_1,...,x_p)$ can be measured on each unit.

Manly et al. (1993, p. 29-30) outline the following assumptions in estimating resource selection functions: (a) the distributions of the measured $\mathbf{x}$ variables for resource units and the resource selection probability function do not change during the study period; (b) the population of available resource units has been correctly identified; (c) the subpopulations of used and unused resource units have been correctly identified; (d) the $\mathbf{x}$ variables that influence the probability of selection have been correctly identified and measured; (e) animals have equal access to all available resource units; and (f) resource units are sampled randomly and independently.

It is important to note that in order to estimate a resource selection probability function, the researcher must first identify the population of available resource units and the significant variables $\mathbf{x} = (x_1,...,x_p)$ that influence the probability of selection for a resource unit. In each study, the researcher must use sound statistical methods to select the influential variables before constructing a function to estimate probability of selection.

2.7 Binary Response Variable

As we have discussed, the data for resource selection studies may be collected through diverse sampling methods, and the variables observed for resource units may be discrete or continuous. However, here we consider selection studies

where the response variable is the use or rejection of a resource unit.

We assume that an animal selects a resource unit based on particular characteristics of that resource unit in comparison with characteristics of other available units. Therefore, an animal responds with a Yes (Y=1) or a No (Y=0) to a particular resource unit configuration, in other words, to specific values of some explanatory variables that characterize the resource unit (Ramsey et al. 1994).

3. Logistic Regression Analysis in Resource Selection Studies

In this section, we review logistic regression as an approach to analyze resource selection studies where the response variable is the use or rejection of a resource unit. We discuss analyses conducted on habitat selection studies of the spotted owl in western Oregon, fernbirds in Otago, New Zealand, and prehistoric Maya communities in Corozal District, Belize, that illustrate some of the issues introduced in the previous section.

3.1 Logistic Regression

Logistic regression is an available procedure to model binary data within the GLM framework. In a resource selection study, let Y represent whether a resource unit is selected, and let $\pi(\mathbf{x})$ be the probability of unit selection, where $\mathbf{x} = (x_1,x_2,...,x_p)$ is a column vector of explanatory variables. Denoting a response with Y=1 if the resource unit is selected and Y=0 if it is rejected yields a Bernoulli random variable with mean $E(Y) = \pi(\mathbf{x})$ and variance to mean relation

$$\text{var}(Y) = \pi(\mathbf{x})\,(1 - \pi(\mathbf{x})). \tag{4}$$

Equation 4 indicates that the variance of Y is not constant, and since Y is binary, it does not have a normal distribution. Therefore, the ordinary least squares estimators from the linear probability model

$$E(Y) = \pi(\mathbf{x}) = \alpha + \beta'\mathbf{x},$$

where $\beta'=(\beta_1,\beta_2,...,\beta_p)$ is a row vector of regression parameters, do not apply, and we expect a nonlinear relationship between $\pi(\mathbf{x})$ and $\mathbf{x}$. The logistic regression function

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta'\mathbf{x})}{1 + \exp(\alpha+\beta'\mathbf{x})}$$

implies a curvilinear relationship between $\mathbf{x}$ and $\pi(\mathbf{x})$. For this model, the odds of a response Y=1 (selection of a resource unit) are

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \exp(\alpha+\beta'\mathbf{x}) = e^{\alpha}(e^{\beta'})^{\mathbf{x}}.$$

Therefore, if $x_2,...,x_p$ are held constant, the odds of selection (Y=1) increase multiplicatively by a factor of e^{β_1} for every unit increase in x_1. The log odds transformation

$$\text{logit}(\pi(\mathbf{x})) = \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \alpha + \beta'\mathbf{x} \tag{5}$$

is the appropriate link function for which the logistic regression model is a GLM. The logit acts as a link between the expectation $\pi(\mathbf{x})$ of Y and the linear predictor $\mathbf{x}$. In addition to the interpretation of the logit link as a log odds ratio providing multiplicative effects, it is important to note that logistic regression provides the same parameters (except for the intercept) under prospective or retrospective studies.

The regression parameters must be estimated in order to fit a logistic regression model to a set of data. Parameters can be estimated by a maximum likelihood method as presented by Kaur et al. (1995). In order to model binary data in a full-likelihood approach, let $\mathbf{y} = (y_1,y_2,...,y_n)$ be a vector of observations for the response $\mathbf{Y} = (Y_1,Y_2,...,Y_n)$ such that $Y_i \approx \text{bin}(1,\pi_i)$. Then the log likelihood in terms of the canonical parameter $\pi = (\pi_1,...,\pi_n)$ is

$$L(\pi;\mathbf{y}) = \sum_{i=1}^{n}\left[\, y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right]. \tag{6}$$

From Equation 2, the logistic regression link function assumes the form

$$\text{logit}(\pi_i) = \eta_i = \sum_{j=1}^{p}\beta_j x_{ij}.$$

Therefore, the log likelihood (Equation 6) is expressed as a function of regression parameters in the form

$$L(\pi(\beta);\mathbf{y}) = \sum_{i=1}^{n}\sum_{j=i}^{p} y_i\, x_{ij}\,\beta_j - \sum_{i=1}^{n}\log\left(1+\exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)\right). \tag{7}$$

Within the GLM framework, deviance is the most common measure of a model's goodness-of-fit to the observed data. Therefore, in a model we wish to include variables that greatly reduce deviance. In order to define it, let $L(\hat{\pi};\mathbf{y})$ be the maximum log likelihood given that a proposed model holds. The maximum achievable log likelihood $L(\tilde{\pi};\mathbf{y})$ occurs for the most general model having as many parameters as observations, so that $\tilde{\pi}_i = y_i$ (Agresti 1990, p. 83). The deviance of a GLM is defined as

$$2[L(\tilde{\pi};\mathbf{y})-L(\hat{\pi};\mathbf{y})]. \tag{8}$$

From Equation 6, in logistic regression the log likelihood can be expressed as

$$L(\hat{\pi};\mathbf{y}) = \sum_{i=1}^{n}[y_i\log\hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]. \tag{9}$$

Therefore, the deviance (Equation 8) becomes

$$2\sum_{i=1}^{n} y_i\log\left(\frac{y_i}{\hat{\pi}_i}\right) - 2\sum_{i=1}^{n}(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right).$$

For some generalized linear models, the deviance has approximately a χ^2 distribution when the model holds; the degrees of freedom are equal to the number of observations minus the number of parameters in the model (Agresti 1990, p. 83). On the hypothesis of no relationship between the odds of animal selection and a particular variable x_i, the reduction in deviance for entering x_i into a logistic regression model approximates a χ^2 distribution with one degree of freedom. The deviance (and reduction of deviance) statistic will be

used here to assess goodness-of-fit for various logistic regression models.

3.2 Resource Selection Probability Function in Logistic Regression

Manly et al. (1993) indicate that it is possible to estimate a resource selection probability function using logistic regression when a study involves the collection of only two samples; for instance, a sample of used resource units and a sample of available units. Based on the following relationship between the Poisson and binomial distributions, we can justify the use of logistic regression with samples of resource units.

Consider the case where we have two samples of resource units. Suppose the observations in the first sample are classified into I different classes, and let Y_{i1} , the count of observations in class i at time t, have a Poisson distribution with mean μ_{i1}. Similarly, suppose the observations in the second sample are classified into I classes at the same time, and let . Then the distribution of Y_{i1}, conditional on $Y_{i1}+Y_{i2}=n_i$ for some value n_i, is a binomial distribution with mean

$$E(Y_{i1} \mid Y_{i1} + Y_{i2} = n_i) = n_i \pi_i \tag{10}$$

and variance

$$\text{var}(Y_{i1} \mid Y_{i1}+Y_{i2} = n_i) = n_i \pi_i (1-\pi_i), \tag{11}$$

where $\pi_i = [\mu_{i1}/(\mu_{i1} + \mu_{i2})]$ is the probability of a type i observation being in the first sample rather than in the second sample, given that it is in one of the two samples (McCullagh and Nelder 1983).

Therefore, if we have two samples of resource units of different types, we can think of π_i as a binomial function of $\mathbf{x}$ variables. Then, the parameters of a resource selection probability function can by estimated by logistic regression. There are three possible scenarios for analysis depending on the types of units in the two samples: (a) a sample of available and a sample of used units at time t, (b) a sample of available and a sample of unused units at time t, and (c) a sample of unused and a sample of used units at time t. These three situations are discussed in the following subsections as presented by Manly et al. (1993).

3.2.1 Samples of Available and Used Resource Units. The population of available resource units (before any unit is used) consists of I classes where the A_i units in class i, for $i=1,...,I$, have vectors of observations $\mathbf{x_i} = (x_{i1},x_{i2},...,x_{ip})$. We consider the situation where this population is sampled so that every available resource unit has the same probability P_a of being included in a random sample of available units. Similarly, every unit that has been used after a period of selection has the same probability P_u of being included in a random sample of used units.

Suppose that there are a_i class i units in the available sample and u_i class i units in the used sample. Also assume that the population of resource units is large and the sampling probabilities are small so that a_i and u_i have approximately independent Poisson distributions. Then the mean of a_i is

$$E(a_i) = P_a A_i \tag{12}$$

and the mean of u_i is

$$E(u_i) = P_u A_i \omega_i (\mathbf{x_i}), \tag{13}$$

where $\omega_i (\mathbf{x_i})$ is the resource selection probability function (the probability of use for a type i unit). We assume that the resource selection probability function takes the form

$$\omega(\mathbf{x}) = \exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p). \tag{14}$$

where the argument of the exponential function should be negative and the selection time is t=1. It follows then from Equations 10 and 11 that u_i has a binomial distribution, conditional on the total number of class i units in both samples being equal to $n_{i=u_i+a_i}$, with mean $n_i\pi_i$ and variance $n_i\pi_i(1-\pi_i)$, where

$$\pi_i = \frac{\exp[\log(P_u/P_a) + \alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip}]}{1 + \exp[\log(P_u/P_a) + \alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip}]}.$$

This is a logistic regression function where α is replaced by a modified parameter $\log (P_u/P_a) + \alpha$ to account for different sampling probabilities for available and used resource units. The dependent variable in the logistic regression is the number of used units of class i out of the total number of sampled units of class i.

Since the constant in the logistic regression is $\log (P_u/P_a) + \alpha$, then the parameter α in the resource selection probability function can be estimated by subtracting $\log (P_u/P_a)$ from the estimated constant in the regression equation. If the ratio of sampling probabilities $\log (P_u/P_a)$ is unknown, then α cannot be estimated, but it is still possible to estimate the resource selection function

$$\omega^*(\mathbf{x}) = \exp(\beta_1 x_1 +...+ \beta_p x_p) \tag{15}$$

and use it to compare resource units.

3.2.2. Samples of Available and Unused Resource Units. Now we consider the case with independent samples of available and unused resource units. The resource selection probability function can be estimated by

$$\omega(\mathbf{x}) = 1 - \exp(\alpha+\beta_1 x_1+...+\beta_p x_p),$$

where the argument of the exponential function should be negative and the selection time is t=1. Equation 12 gives the mean number of class i units in the sample of available resource units, while the expected frequency $\overline{u_i}$ of class i units in the sample of unused units is given by

$$E(\overline{u_i}) = P_{\overline{u}} A_i \exp(\alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip}), \tag{16}$$

where $P_{\overline{u}}$ is the sampling probability for unused resource units. Assume that these sample counts have independent Poisson distributions. Then by Equations 10 and 11, the distribution of $\overline{u_i}$, conditional on $n_1 = a_i + \overline{u_i}$, is binomial with mean $n_i\pi_i$ and variance $n_i \pi_i (1-\pi_i)$, where

$$\pi_i = \frac{\exp[\log(P_{\overline{u}}/P_a) + \alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip}]}{1 + \exp[\log(P_{\overline{u}}/P_a) + \alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip}]}.$$

This is a logistic regression model where the dependent variable is the number of type i units in the sample of unused

resource units. Again, α can be estimated only if the ratio of sampling probabilities log $(P_{\bar{u}}/P_a)$ is known.

3.2.3 Samples of Unused and Used Resource Units. Now we discuss the case with two independent samples of unused and used resource selection units. In this situation, the expected number of class i units in the sample of unused units is

$$E(\bar{u_i}) = P_{\bar{u}} A_i (1-\omega(\mathbf{x_i})),$$

and the expected number of type i units in the sample of used units is

$$E(u_i) = P_u A_i \omega(\mathbf{x_i}),$$

where $P_{\bar{u}}$ and P_u are sampling probabilities, A_i is the number of available class i units, and $omega(\mathbf{x_i})$ is the resource selection probability function as defined in Equation 14. By Equations 10 and 11, u_i has a binomial distribution, conditional on $n_{i=u_i+\bar{u_i}}$ being the total number of class i units in both samples, with mean $n_i\pi_i$ and variance $n_i\pi_i(1-\pi_i)$, where

$$\pi_i = \frac{P_u A_i \omega(\mathbf{x_i})}{P_{\bar{u}} A_i(1-\omega(\mathbf{x_i})) + P_u A_i \omega(\mathbf{x_i})}$$
$$= \left(\frac{(P_u/P_{\bar{u}})\omega(\mathbf{x_i})}{1-\omega(\mathbf{x_i})}\right)\Big/\left(1+\left(\frac{(P_u/P_{\bar{u}})\omega(\mathbf{x_i})}{1-\omega(\mathbf{x_i})}\right)\right).$$

This equation defines a logistic regression by letting

$$\frac{(P_u/P_{\bar{u}})\omega(\mathbf{x_i})}{1-\omega(\mathbf{x_i})} = \exp(\alpha+\beta_1 x_{i1}+...+\beta_p x_{ip})$$

so that

$$\pi_i = \frac{exp(\alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip})}{1 + exp(\alpha + \beta_1 x_{i1} +...+ \beta_p x_{ip})} \qquad (17)$$

and

$$\omega(\mathbf{x_i}) = \frac{\exp[\log(P_{\bar{u}}/P_u)+\alpha+\beta_1 x_{i1}+...+\beta_p x_{ip}]}{1+\exp[\log(P_{\bar{u}}/P_u)+\alpha+\beta_1 x_{i1}+...+\beta_p x_{ip}]}. \qquad (18)$$

The estimates of the parameters in Equation 17 can be obtained by fitting a logistic regression model where the dependent variable is the number u_i of used units out of the total number of type i units. Then, using the estimates from Equation 17 and if the ratio $(P_{\bar{u}}/P_u)$ of sampling probabilities is known, we can use Equation 18 to estimate the resource selection probability function. If the sampling probability ratio is unknown, the resource selection probability function cannot be estimated. In this case, we can set $(P_{\bar{u}}/P_u)=1$ in Equation 18. The resulting estimated function is an index of selectivity that ranks resource units in the same order they would be ranked if their selection probabilities could be estimated.

3.3 Habitat Association Study of the Northern Spotted Owl

The endangered nothern spotted owl *(Strix occidentalis caurina)* is at the center of controversy since its habitat is associated with the commercially valuable old forests of the Pacific Northwest. Specifically, the owl's range includes the timber rich mountains from northwestern California to southwestern British Columbia. The northern spotted owls are territorial birds and have been shown to select nest sites with significantly higher proportion of old forest than randomly available(Ripple et al. 1991, Ripple et al. 1997). The current debate now revolves around the questions of how dependent are the owls on old forest habitat and how much fragmentation can the owls tolerate.

Ramsey et al. (1994) conducted a resource selection study for the spotted owl in the Pacific Northwest state of Oregon. More specifically, it was a habitat association study that attempted to determine the importance of old forest (80 years or older) in nest site selection for the spotted owl. The following is a discussion of our attempt to build upon the work conducted by Ramsey et al. (1994) using logistic regression techniques.

3.3. 1. Sampling Design and Data. A specific design for habitat association studies was employed by Ripple et al. (1991) in order to collect data. Owl pairs were located using taped vocalizations and whistles in a 2500 km^2 region. Researchers located 37 nest sites by following owls that responded and 30 of the 37 nest sites were randomly selected for study. Thirty additional sites were also selected at random coordinates to act as control sites and compare conditions at used and available sites. Aerial photography was used to determine the percentages of old forest in seven concentric circles around each nest and random site.

We assume that owls respond with a Yes (Y=1) or a No (Y=0) to nesting at a particular site based on its habitat configuration. Notice that the study design fixes the sample size for yes and no responses, so it has the same structure as case-control retrospective studies. As a result, logistic regression becomes available for analysis since it ties the retrospective study to a model for the prospective selection process (Ramsey et al. 1994).

The data for each of the 60 sites consists of seven variable measurements that correspond to the percentage of old forest in seven circles around the site.

3.3.2. Questions of Interest and Statistical Analysis. Ripple et al. (1991) concluded that the percentages of old forest adjacent to a nest and in the surrounding area are important in selecting a nest site. Ramsey et al. (1994) addressed the further issue of how much old forest fragmentation can the spotted owl tolerate for nesting. If spotted owls can tolerate some fragmentation, then timber in their habitat could be managed without threatening the survival of the species.

In order to investigate habitat fragmentation, the data are transformed into percentage of old forest in seven concentric, non-overlapping rings (R_i, i=1,...,7) around a site. Notice that previously the data consisted of old forest percentages in seven concentric circles, but the larger circles contained the smaller ones. Therefore, the analysis of these data provided an idea of the extent to which surrounding old forest is important for nesting. In contrast, analyzing the data for non-overlapping rings may give an idea of the permissible habitat fragmentation for nesting. A sample of the data is given on Table 1, where variable names correspond to outer radii for the rings. Figure 1 shows a clear difference between the distributions of ring percentage of old forest around nest and random sites.

Table 1. Percentages of old forest in concentric rings around spotted owl nest sites (N) and random sites (R) in western Oregon (Source: Ramsey et al. 1994).

Site Type	0.91km	1.18km	1.40km	1.60km	1.77km	2.41km	3.38km
R	26.0	33.3	25.6	19.1	31.4	24.8	17.9
R	100.0	92.7	90.1	72.8	51.9	50.6	41.5
R	32.0	22.2	38.3	39.9	22.1	20.2	38.2
N	80.0	87.3	93.3	81.6	85.0	82.8	63.6
N	96.0	74.0	76.7	66.2	69.1	84.5	52.5
N	82.0	79.6	91.3	70.7	75.6	73.5	66.8

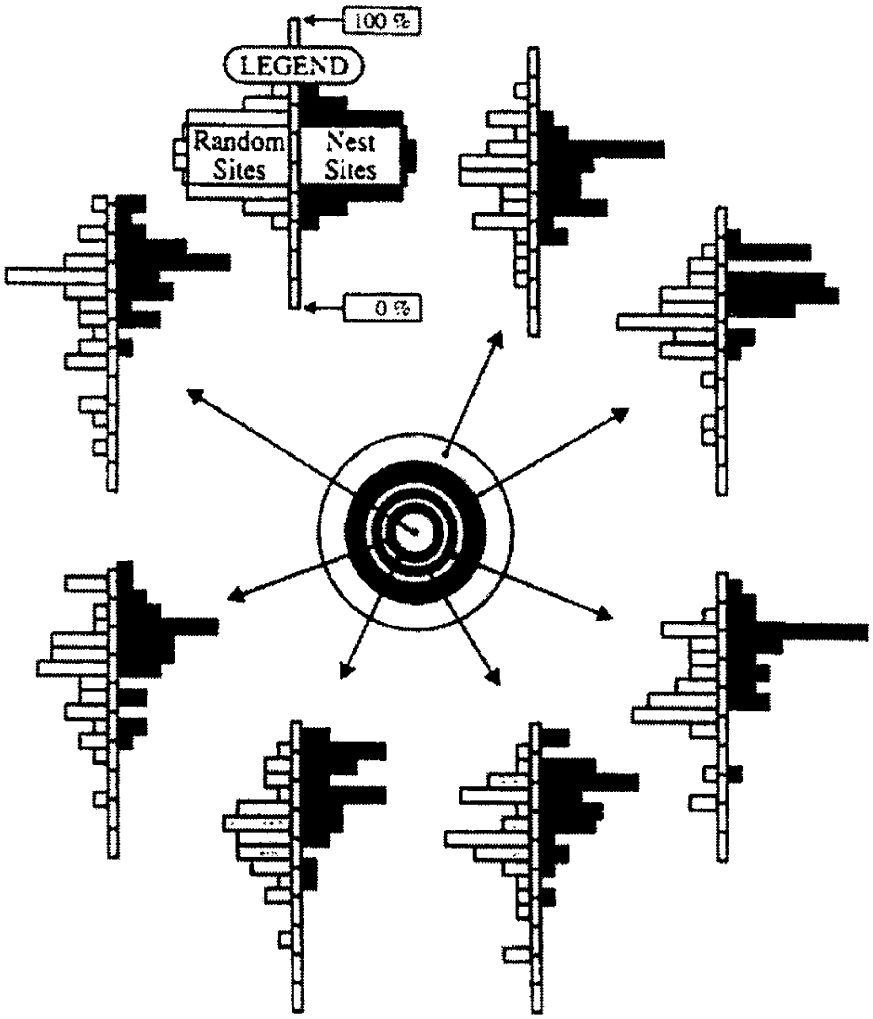

Figure 1. Percentages of old forest in rings surrounding spotted owl nest sites and random sites in western Oregon, 1987-1989 (Source: Ramsey et al. 1994).

Table 2. Habitat Fragmentation Model for Northern Spotted Owl Nesting (Proposed by Ramsey et al. 1994).

Coefficients	Value	Std. Error	t value
Intercept	-0.966676872	9.674590593	-0.09991915
R_1	0.630204513	0.382458274	1.64777325
R_2	0.505918554	0.353933147	1.42941840
R_3	-0.785145053	0.482729481	-1.62647007
R_5	-0.567139585	0.259488888	-2.18560258
R_1^2	-0.008481794	0.003925561	-2.16065745
R_2^2	0.012638775	0.006391750	1.97735750
R_3^2	0.016135873	0.008021850	2.01149029
$R_1 * R_3$	0.009810176	0.005191387	1.88970230
$R_2 * R_3$	-0.036620167	0.016809937	-2.17848322
$R_3 * R_5$	0.009900098	0.004186713	2.36464709

Table 3. Analysis of Deviance for Habitat Fragmentation Model (Terms added sequentially: first to last).

Term	Df	Dev. Reduction	Model Df	Deviance
Null			59	83.17
R_1	1	12.07	58	71.10
R_2	1	0.03	57	71.07
R_3	1	11.16	56	59.91
R_5	1	3.90	55	56.01
R_1^2	1	0.59	54	55.42
R_2^2	1	0.28	53	55.14
R_3^2	1	0.002	52	55.14
$R_1 * R_3$	1	1.52	51	53.62
$R_2 * R_3$	1	3.99	50	49.63
$R_3 * R_5$	1	8.56	49	41.07

An association between nest site selection and habitat fragmentation may be apparent in the data in three ways (Ramsey et al. 1994): (a) as curvature if owls select sites with more than average available old forest but less than the maximum percentage available; (b) as interactions between percentage of old forest in different rings – for example, a high percentage of old forest in R_1 may produce a tolerance for low percentage in R_2 yielding the interaction between both rings significant; (c) as negative coefficients if there is a preferred fragmentation pattern corresponding to rings chosen by the researchers, but this is unlikely to happen. These effects would appear respectively as quadratic terms, interactions, and negative coefficients in logistic regression models.

Therefore, to address the possibilities for curvature and interactions, Ramsey et al. started with a logistic regression model that included the explanatory variables R_1,R_2,R_3,R_5 that correspond to percentage of old forest surrounding a site in rings 1,2,3, and 5, all their two-way interactions, and their quadratic terms. Note that rings 4 and 6 were excluded from the original model since their main effect coefficients were less than their standard errors in a most general quadratic logit expression. Following a stepwise logistic regression procedure, Ramsey et al. obtained the final model that is reproduced and summarized on Table 2. An analysis of deviance is summarized on Table 3.

Discussion

In view of the three general sampling designs discussed in section 2.4, the sampling procedures for the spotted owl study fall generally within design I. The resource units are potential nesting sites; available resource units (random sites) were sampled for the population of animals in the entire study region. The researchers used sampling protocol A since they sampled random available sites and nest (used) sites.

Recall that the sample sizes for observing used and available sites were fixed at thirty, so the study design corresponds to a retrospective case-control study. Therefore, it is an observational study, and no cause-effect relationship can be inferred. However, Ramsey et al. indicate that the case-control design is helpful in determining the direction of future research.

With regard to the statistical analysis presented here, no definite evidence was found for a preference for habitat fragmentation. The final model included some quadratic and in-

teraction terms – suggestive evidence of curvature and interaction among percentages of old forest in rings surrounding sites. However, Ramsey et al. warn that this evidence is inconclusive; furthermore, he argues that although these features may be related to a tolerance for some habitat fragmentation, other explanations are possible.

Furthermore, some of the variables in the final model are hard to interpret. It is important to note that the home range radius for this area has a median of 2.17 km, a mean of 2.37 km, and a standard deviation of 0.30 km (Keith Swindle, personal communication 1996). In the model, the area immediately surrounding a site is important for nesting; hence the significance of the inner rings R_1, R_2, and R_3 that combine for a radius of 1.40 km around a site. The outer ring R_7, with a 2.41 km to 3.38 km radius around the site, does not appear to be important for nest selection since it is just outside the average home range radius for this area. The significance of R_5 may indicate some tolerance for old forest fragmentation. The interactions $R_1{*}R_3$, $R_2{*}R_3$ and R_{3*R_5} indicate that the combination of old forest patterns in those rings are important for nesting. A possible interpretation is that owls tolerate a lower proportion of old forest in the third ring as long as there is a high proportion of old forest in the first and second rings. Hence the significance of the $R_1{*}R_3$ and $R_2{*}R_3$ interactions. This interpretation is only a conjecture, however. At best, the interactions may indicate that some habitat fragmentation is permissible for nesting.

With regard to the analysis of deviance on Table 3, recall that, for each term, deviance reduction has a χ^2 distribution with 1 degree of freedom. Therefore, we can observe that the main effects and the interaction effects significantly reduce deviance in the logistic regression model; in contrast, the quadratic terms do not contribute significantly to improve the model's goodness-of-fit in the presence of main effects. The meaning and importance of the quadratic effects are dubious.

The difficulty in interpreting results with the available information indicates, as stated by Ripple et al. (1991), that additional data are necessary to investigate conclusively the issue of habitat fragmentation.

3.3.3 Analysis with Modified Rings. In section 2.4, we saw that the choice of study area and its boundaries has an important impact on subsequent analysis. Although the data for this spotted owl study has been collected for a determined radius around sites, it may be informative to explore the habitat fragmentation question by modifying the arbitrary seven rings. Combining the data for R_1, R_2, and R_3 into one inner ring R_1^*; R_4 and R_5 into a middle ring R_2^*; and R_6 and R_7 into an outer ring R_3^* may provide a different exploratory insight on the habitat fragmentation problem. A sample of the new data is shown on Table 4; variable names correspond to outer radii of modified rings.

We start with a logistic regression model including as explanatory variables the modified rings R^{*1}, R^{*2}, R^{*3}, their two-way interactions, and their quadratic terms. The final model, reached through a stepwise regression procedure, is summarized on Table 5. An analysis of deviance is given on Table 6.

Examining the new model, we observe that the arbitrary choice of ring area has some impact on the analysis and its interpretation. The outer ring R^{*3} is now significant; this indicates that the percentage of old forest between a 1.77 km and a 3.38 km radius is important for nest site selection. The significant interaction $R^{*1}{*}R^{*3}$, and the quadratic terms , however, may reaffirm the evidence for some habitat fragmentation tolerance as it appeared in the original analysis by Ramsey et al. The analysis of deviance shows that the inner ring (1.40 km radius around a site) and the interaction of the inner and outer rings have the most significant impact in improving the fit of the model. The difficulty in drawing conclusions from the available data remains, but we see that the choice of ring sizes influences the results of the analysis.

Table 4. Percentages of old forest in modified concentric rings around spotted owl nest sites (N) and random sites (R) in western Oregon.

Site Type	1.40km	1.77km	3.38km
R	27.98	25.11	20.12
R	95.03	65.59	44.44
R	31.00	31.21	32.39
N	83.98	86.89	53.30
N	85.95	83.26	69.79
N	84.08	67.62	62.82

Table 5. Modified Rings Model for Habitat Fragmentation.

Coefficients	Value	Std. Error	t value
Intercept	-2.672497471	6.458445604	-0.4137989
R_1^*	0.033897915	0.218809639	0.1549197
R_2^*	0.064203956	0.044869823	1.4308939
R_3^*	-0.106465795	0.267220215	-0.3984197
R_1^{*2}	-0.003313715	0.002506..51	-1.3218109
R_3^{*2}	-0.004347605	0.002777174	-1.5654782
$R_1^* * R_3^*$	0.008282558	0.004150467	1.9955727

Table 6. Analysis of Deviance for Modified Rings Model (Terms added sequentially).

Term	Df	Dev. Reduction	Model Df	Deviance
Null			59	83.18
R_1^*	1	13.59	58	69.59
R_2^*	1	1.70	57	67.89
R_3^*	1	0.47	56	67.42
R_1^{*2}	1	0.08	55	67.34
R_3^{*2}	1	0.65	54	66.69
$R_1^* * R_3^*$	1	6.79	53	59.90

3.3.4 Resource Selection Function for Habitat Fragmentation. In order to estimate a resource selection probability function for the spotted owl study, we need to assess the assumptions given in section 2.6. It is reasonable to assume that there is a population of N available sites for nesting. For each site, we can measure the variables $\mathbf{R} = (R_1,...,R_7)$ corresponding to percentages of old forest in seven rings around each site. It was assumed that the percentages of old forest remained the same during the study period, so the distributions of the R_i variables did not change. We must assume that the researchers correctly identified the population of available nest sites and the subpopulation of used nest sites. It is reasonable to believe that owls select the best available sites for nesting. However, the presence of other territorial owls would limit the access to some sites. In order to assess the model proposed by Ramsey et al., it is assumed here that logistic regression is a sound approach to identify the significant variables that influence the probability of selecting a site for nesting.

The final assumption is that resource units are sampled randomly and independently. For the spotted owl study, researchers used sampling protocol A. Therefore, data consists of a sample of used (nest) sites and a sample of available sites. It is reasonable to assume that these samples are independent and that available sites were randomly sampled since the researchers chose sites at random coordinates. However, it would be difficult to justify that the sample of nest sites is random as only 37 owl nests were located and 30 of them were selected for the study. Manly et al. (1993) indicate that when resource units are not sampled randomly, the estimates of the coefficients of a resource selection function may still be meaningful but standard errors may not reflect true variation in the populations. Therefore, caution must be exercised in estimating a selection function for the spotted owl study.

Since we have a sample of nest (used) sites and a sample of available sites, Equation 14 provides the form of the resource selection probability function for owl nesting. In order to estimate this function, however, every nest site must have the same probability P_u of being sampled. We already determined that this is unlikely since only 37 nests were located and the sampling method (listening for owls that responded to whistles and taped vocalizations) could have favored locating owls nesting near forest clearings. Furthermore, the sampling probabilities ratio (P_u/P_a) is unknown, so the nest selection probability function cannot be estimated.

However, it is still possible to estimate a resource selection function as given in Equation 15. This function is a selectivity index that ranks sites in the order of their probabilities of selection for owl nesting. Recall the model proposed by Ramsey et al. for habitat fragmentation (Table 2). The parameter estimates obtained through logistic regression yield the nest site selection function

$$\omega^*(\mathbf{R})=\exp[0.63R_1+0.51R_2-0.79R_3-0.57R_5-0.01R_1^2+...$$
$$...+0.01(R_3*R_5)].$$

This function may be used to compare nest sites. Ignoring the quadratic and interaction terms, the positive coefficients for R_1 and R_2 may indicate that owls prefer high densities of old forest near nest sites, while the negative coefficients for R_3 and R_5 indicate tolerance for low densities of old forest in outer rings.

We can also estimate a resource selection function based on the parameters from the modified rings model (Table 5) for habitat fragmentation to obtain

$$\omega^*(\mathbf{R})=\exp[0.033R_1^*+0.064R_2^*-0.106R_3^*-0.003R_1^{*2}-$$
$$...+0.008(R_1^**R_3^*)].$$

Again the positive coefficients for R^{*1} and R^{*2} and the negative coefficient for R^{*3} may indicate a tolerance for habitat fragmentation in the outer rings. In fact, if we fit a main effects logistic regression model for the modified rings data, we obtain the model shown on Table 7. Based on this simplified model, the resource selection function for habitat fragmentation would be

$$\omega^*(\mathbf{R})=\exp[0.042R_1^*+0.058R_2^*-0.024R_3^*]$$

so that, for nesting, owls prefer higher proportion of old forest near the nest sites and tolerate lower density in the outer ring.

The concept of a resource selection function is useful in providing more information on the issue of habitat fragmentation. A selectivity index is now available to rank resource units based on the probability of owl nesting for different fragmentation patterns; furthermore, we have found some suggestive evidence that owls select nest sites in areas with some forest fragmentation outside a 1.77 km radius. We must further refer the reader to the study by Boyce et al. (1994),

Table 7. Main Effects Rings Model for Habitat Fragmentation.

Coefficients	Value	Std. Error	t value
Intercept	-5.39414922	2.11701671	-2.5479956
R_1^*	0.04233438	0.03329770	1.2713904
R_2^*	0.05752710	0.04081728	1.4093809
R_3^*	-0.02415681	0.03569472	-0.6767616

where resource selection probability functions are also used in the study of spotted owl habitat.

3.3.5 Extended Analysis with Redefined Fragmentation Variables. The concept of ring variables is artificial; rings were defined in an attempt to investigate habitat fragmentation with the available data. In a similar manner, we now define three new variables in order to construct a simple model for the spotted owl study. We suggested in the previous section that owls tolerate forest fragmentation in the outer rings R_6 and R_7. With redefined variables, we now investigate habitat fragmentation only within a 1.77 km radius around sites.

Let x_1 be the percentage of old forest within a $1.77km$ radius from a site; this variable represents the total area of old forest immediately surrounding a site since it combines the first five rings, R_1 through R_5. Let x_2, the minimum percentage of old forest in any of the first five rings, represent an owl's maximum tolerance for old forest depletion in a ring surrounding a nest site. Finally, let x_3 be the difference between the maximum and the minimum percentage of old forest in any of the first five rings; then, x_3 represents the range in the density of old forest that an owl tolerates around a nest. A sample of the new data is presented on Table 8. A model including these three variables may contribute to our understanding of habitat fragmentation.

We fit a main effects logistic regression model for the redefined variables. The results of the model and the corresponding analysis of deviance are presented on Tables 9 and 10, respectively. The model seems appropriate for the data; our set of variables significantly reduces deviance. In previous analyses, we had already determined that the percentage of old forest around a site is important for nest selection; owls clearly prefer forested areas. It is more informative

Table 8. Redefined fragmentation variables: percentage of old forest within a 1.77 km radius around spotted owl nest and random sites.

Site Type	Total	Minimum	Range
R	26.90	19.10	14.20
R	82.89	51.90	48.10
R	31.08	43.00	17.80
N	85.07	80.60	12.30
N	84.95	80.00	13.30
N	77.91	66.20	29.80

Table 9. Redefined Variables: Habitat Fragmentation Model for Northern Spotted Owl Nesting.

Coefficients	Value	Std. Error	t value
Intercept	-3.94264742	2.21852237	-1.7771502
x_1	0.09564732	0.13381547	0.7147703
x_2	-0.01791405	0.13144678	-0.1362837
x_3	-0.07593159	0.09603438	-0.7906709

Table 10. Analysis of Deviance for Fragmentation Model (Terms added sequentially).

Term	Df	Dev. Reduction	Model Df	Deviance
Null			59	83.18
x_1	1	15.21	58	67.97
x_2	1	2.56	57	65.41
x_3	1	0.63	56	64.78

to discuss the meaning of x_2 and x_3. A resource selection function based on our logistic regression model is

$$\omega^*(\mathbf{x}) = \exp(0.096x_1 - 0.018x_2 - 0.076x_3).$$

As expected, a higher total percentage of old forest increases the probability of nesting in a site. A relatively low percentage of old forest in one of the inner rings slightly increases the probability of nesting. It is plausible that moderate fragmentation is tolerated. However, a large difference between the maximum and the minimum percentage of old forest in the innermost five rings around a site diminishes the probability of owl nesting. This suggests that owls prefer forest uniformity; nest selection occurs mostly in forested areas without large forest cuttings or clearings.

3.4 Nest Selection Study of Fernbirds

Another habitat selection study where the response variable is the use or rejection of a resource unit is conducted by Harris (1986) on nest selection by fernbirds in Otago, New Zealand.

3.4.1. Sampling Design and Data. Harris (1986) found 24 fernbird nest sites in 1982-1983 and 1983-1984 and measured nine variables on each site. He sampled 25 comparable sites by randomly choosing points in the study area. The study has sampling design I since potential nest sites were sampled for the entire study area and for all fernbirds in

Table 11. Data for Fernbird Nest Selection Study (Source: Manly et al. 1993).

Site Type	Canopy height (m)	Distance to edge (m)	Perimeter of clump (m)
R	0.47	13.5	3.17
R	0.62	8.0	3.23
R	0.75	19.0	2.44
N	1.20	14.0	8.90
N	0.58	25.0	4.34
N	0.74	14.0	2.30

Table 12. Model for Fernbird Nest Selection Study.

Coefficients	Value	Std. Error	t value
Intercept	-10.7279571	3.2986814	-3.252196
Height	7.7957204	3.2429880	2.403870
Distance	0.2097201	0.1208261	1.735719
Perimeter	0.8836828	0.4792432	1.843913

the area; the samples of used and available resource units correspond to sampling protocol A. Harris concluded that only three variables were significant in nest selection. These variables were the canopy height, the distance from the outer edge of the nest to the nearest outer surface of the clump of vegetation where the nest is located, and the perimeter of the clump of vegetation (Manly et al. 1993). A sample of the data is shown on Table 11.

3.4.2 Statistical Analysis. Given the nature of the response variable (Y=1 for a nest site, *Y*=0 for an available site), logistic regression is appropriate for analysis within the GLM framework. Logistic regression in Splus provides the model on Table 12 for the fernbird study. Table 13 presents an analysis of deviance. It is clear that all three variables significantly improve the fit of the model by reducing deviance. The model indicates that surrounding vegetation is important for fernbird nest selection.

3.4.3 Resource Selection Function for Fernbird Study. Manly et al. (1993) estimated a resource selection function based on the logistic regression model on Table 12. There is a sample of nest (used) sites and a sample of available sites; we assume the probabilities of sampling an used site (P_u) and an available site (P_a) are both small. However, these probabilities cannot be estimated; hence, the sampling probabilities ratio (P_u/P_a) is unknown. In this case, Equation 15 yields the resource selection function

$$\omega^*(x)=\exp[7.80(Height)+0.21(Distance)+0.88(Perimeter)]$$

that provides a selectivity index for fernbird nesting. Higher canopy is the characteristic that most improves the probability of nesting in a particular site.

3.5 Selection of Prehistoric Maya Settlements

The study of the use of resources by human societies is similar to the study of resource selection by animals since human societies are selective in choosing resources among available options. Green (1973) conducted a study on location of prehistoric Maya sites in Corozal District, Belize. The study evaluates the proposition that settlements were selected in order minimize the necessary effort to acquire critical resources. Green attempts to identify which environmental characteristics were relevant in selecting a place for settlement. A related goal is to predict the location of sites in portions of the study region that had not yet been explored archaeologically at the time of the study. Here we analyze these issues using logistic regression.

3.5.1 Sampling Design and Data. The study region was divided into 151 square plots of land with 2.5 km sides; these 151 plots are the resource units. For each plot, thirteen variables related to the natural and social environment were measured. The response variable considered by Green (1973) is the number of Maya sites in a plot of land – 24 plots had one site and 5 plots had two sites giving a total of thirty-four sites in 29 plots.

Green observes that sites could not be sampled randomly from the universe of all prehistoric sites; selected sites rather represent those that could be accurately located. Although the sample is as complete as possible at the time of the study, it does not include all prehistoric sites in the region since lower, smaller structures are destroyed by commercial farming and local settlement. Thus, the sample is probably biased towards denser, larger settlements and omits some small individual residence clusters.

Most of the thirteen variables considered for analysis pertain to the natural environment even though some refer to the social environment. The variables are related to soil types, vegetation types, distance from navigable water, distance from Santa Rita (a prehistoric commercial center), and number of sites in surrounding plots of land. For analysis, variables are defined as follows: x_1 is the percentage of soils under constant lime enrichment; x_2 is the percentage of meadow soil with calcium ground water; x_3 is the percentage

Table 13. Analysis of Deviance for Fernbird Model (Terms added sequentially).

Term	Df	Dev. Reduction	Model Df	Deviance
Null			48	67.93
Height	1	17.11	47	50.82
Distance	1	6.36	46	44.46
Perimeter	1	3.98	45	40.48

Table 14. Multiple Regression Model for Prehistoric Maya Sites.

Coefficients	Value	Std. Error	t value	p value
Intercept	0.2644	0.1535	1.7225	0.0872
Lime-Soil	0.0030	0.0017	1.8019	0.0738
Meadow-Soil	0.0024	0.0020	1.1894	0.2363
Coral-Bedrock	0.0015	0.0018	0.8158	0.4160
Alluvial-Soil	0.0032	0.0017	1.8970	0.0599
Broadleaf-Forest	-0.0030	0.0020	-1.4859	0.1396
Marsh-Forest	-0.0058	0.0019	-3.0227	0.0030
Palm-Forest	-0.0028	0.0045	-0.6294	0.5301
Mixed-Forest	-0.0009	0.0029	-0.3126	0.7550
Soil Boundaries	0.0208	0.0294	0.7084	0.4799
Distance to Water	-0.0018	0.0165	-0.1086	0.9137
Percentage near Water	0.0003	0.0009	0.3483	0.7281
Santa Rita	-0.0041	0.0035	-1.1752	0.2419
Neighboring Sites	0.0555	0.0256	2.1652	0.0321

of soils formed from coral bedrock under constant lime enrichment; x_4 is the percentage of alluvial and saline organic soils adjacent to rivers and the coast; x_5 is the percentage of deciduous seasonal broadleaf forest; x_6 is the percentage of high and low marsh forest, herbaceous marsh and swamp; x_7 is the percentage of cohune palm forest; x_8 is the percentage of mixed forest of the types listed in x_5 and is the number of soil boundaries in the plot; x_{10} is the distance (km) to navigable water; x_{11} is the percentage of the plot within $1km$ of navigable water; x_{12} is the distance (\$km\$) from Santa Rita; and x_{13} is the number of sites immediately surrounding squares.

3.5.2 Statistical Analysis. In order to identify the significant variables that influence settlement location, Green used a multiple regression model with all thirteen explanatory variables; the response variable was the number of sites in each plot of land (entered as the natural logarithm of 1 + the number of sites). No stepwise procedures or goodness-of-fit measures were considered; the model is reproduced on Table 14.

The significant variables were x_1, x_4, x_5, x_6, and x_{13}. By measuring a simple correlation coefficient between each of these variables and the response, Green arrived at the following conclusions: settlements were selected at places with great expanse of lime enriched soil (fertile for agriculture); alluvial and organic soils adjacent to rivers and the coast were avoided (danger of floods and pests); deciduous seasonal broadleaf vegetation (an indication of fertile soil for agriculture) was preferred; marsh forest and swamp were avoided (poor agricultural potential); and settlements tended to cluster. Based on simple correlation, she also ventured that proximity to navigable water was important because of commercial reasons. According to Green, then, natural habitat selection was guided by good agricultural land and proximity to water trade routes. Finally, she proposed that a multiple regression model with all thirteen explanatory variables be used to predict location of undiscovered sites.

The response variable considered by Green, however, only takes three values (log(1), log(2), log(3)) in the observed data. Therefore, the assumption in multiple regression of normal distribution for the response variable is suspicious. In fact, a normal probability plot (Figure 2) suggests that the errors are not normally distributed since the plot is not a straight line but a curve. Instead of applying a transformation

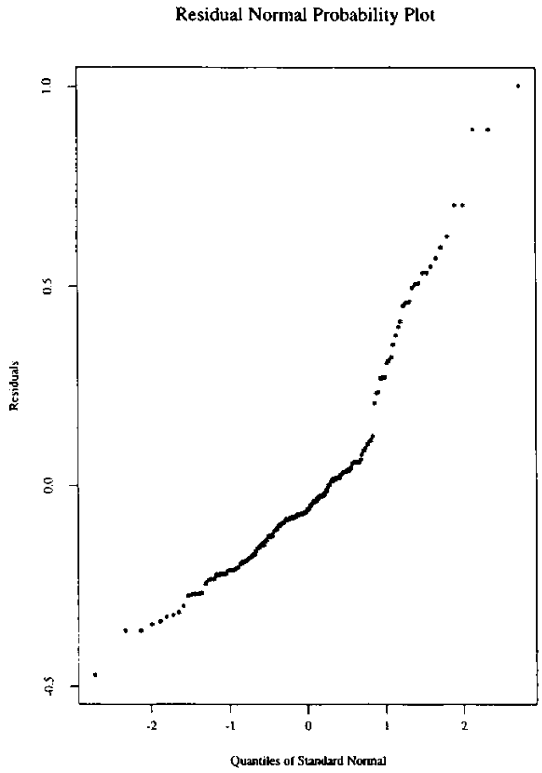

Figure 2. Multiple Regression Residuals for Prehistoric Maya Sites.

Table 15. Model for Selection of Prehistoric Maya Settlements.

Coefficients	Value	Std. Error	t value
Intercept	-2.44627689	0.834572811	-2.931173
Lime-Soil	0.01510416	0.008635984	1.748979
Alluvial-Soil	0.01795122	0.012640751	1.420107
Marsh-Forest	-0.02838368	0.009413119	-3.015332
Neighboring Sites	0.68648031	0.220102266	3.118915

Table 16. Analysis of Deviance for Settlements Model (Terms added sequentially)

Term	Df	Dev. Reduction	Model Df	Deviance
Null			150	147.73
Lime-Soil	1	5.32	149	142.41
Alluvial-Soil	1	0.39	148	142.02
Marsh-Forest	1	16.51	147	125.51
Neighboring Sites	1	10.37	146	115.14

$g(\mathbf{Y})$ to the response, we take advantage of the distributional flexibility under GLM and redefine the response variable to be the presence (Y=1) or absence (Y=0) of sites in a plot of land. Furthermore, to reach a model more parsimonious than Green's, we use a stepwise logistic regression procedure in Splus; the initial model includes all thirteen explanatory variables. The final model and an analysis of deviance are presented on Tables 15 and 16 respectively.

The stepwise procedure identifies the significant variables , x_6, and x_{13} related to settlement location. The analysis of deviance shows that the variables corresponding to the percentage of lime-enriched soil (x_1), the percentage of marsh forest and swamp (x_4), and the number of sites immediately surrounding squares (x_{13}) greatly improve the fit of the model. Types of soil and vegetation, and the presence of neighboring sites are associated with settlement in a particular plot. We cannot dispute the archaeological reasons to believe that distance to navigable water was important for settlement location; however, we can dispute the claim that the data support this belief. The proximity of navigable water, represented by variables x_{10} and x_{11} in the data set, is not significantly associated with settlement. In fact, Green based her claim solely on a positive simple correlation coefficient r=0.09158 between x_{11} and the original response – number of sites in a plot of land. This clearly is not sufficient evidence of a significant association between proximity of navigable water and settlement.

In general, then, the data show that fertile soil for agriculture and the presence of neighboring sites were important in prehistoric Maya habitat selection; however, the data do not support the argument that accessibility of water routes for trading was a significant consideration for settlement.

3.5.3 Resource Selection Function for Prehistoric Maya Settlements. An important corollary goal in the location analysis of prehistoric Maya sites is to predict the location of undiscovered sites in the study region. Since the inadequacy of a multiple regression approach has been suggested, a resource selection probability function based on logistic regression is a more effective tool to identify the plots of land where archaeological excavation has higher probability of success.

Under this approach, however, we must acknowledge the presence of some misclassification since undiscovered sites may exist at plots of land that have been recorded as unused. Therefore, the estimated probability of a plot being used is actually multiplied by the probability of a site being discovered in the plot. This should cause no concern under the assumption that, for all existing sites, the probability of a site being discovered is approximately constant. This is a reasonable assumption except for smaller structures that were not preserved in the past.

Note that a census (and not random sampling) of the entire study region was conducted since all used and available resource units (land plots) were examined. Therefore, the estimated resource selection probability function (based on the model on Table 15) is

$$\omega(\mathbf{x})=\pi \bullet P_d=\left[\frac{\exp(-2.446+0.015x_1+0.018x_4-0.028x_6+0.686x_{13})}{1+\exp(-2.446+0.015x_1+0.018x_4-0.028x_6+0.686x_{13})}\right]P_d,$$

where π is the probability that at least one site is located in a plot, and P_d is the probability of a site being discovered. Alternatively, since P_d is assumed to be constant but is not known, an estimated resource selection function from Equation 15 is

$$\omega^*(\mathbf{x})=\exp(0.015x_1+0.018x_4-0.028x_6+0.686x_{13}).$$

Both suggested equations provide a selectivity index for plots of land. Highly ranked plots where no sites have been discovered are prime locations for exploration.

The resource selection function indicates that presence of neighboring sites is the characteristic that most increases the probability of a plot being used for settlement. Preference for lime-enriched soil and avoidance of marsh forest is also suggested. However, the function does not suggest that alluvial

Table 17. Second Model for Selection of Prehistoric Maya Settlements.

Coefficients	Value	Std. Error	t value
Intercept	-1.51714933	1.59522012	-0.9510596
Lime-Soil	-0.05043594	0.03004583	-1.6786335
Percentage near water	0.04260815	0.02097864	2.0310252

Table 18. Analysis of Deviance for Second Maya Model (Terms added sequentially).

Term	Df	Dev. Reduction	Model Df	Deviance
Null			29	40.20
Lime-Soil	1	17.63	27	22.57
Percentage near water	1	6.95	26	15.62

and saline organic soils near rivers and the coast reduce the probability of settlement as Green believes. This discrepancy demands further archaeological consideration. In general, the resource selection function is most useful in helping to identify the best locations for further exploration.

3.5.4 Extended Analysis of Settlement Sites. For our preceding logistic regression analysis, we defined a binary response variable Y, so that Y=1 indicates the presence of at least one Maya site in a plot of land. Recall, however, that in 5 of the 151 plots examined, there were two Maya sites, and in 24 plots there was only one site. Therefore, some loss of information was involved in defining a binary response since plots with one or two sites were treated equally. To account for those plots with two sites, we can perform a second stage analysis.

We define a Bernoulli random variable Z conditional on Y. We let Z=1 when there is more than one site in a plot of land given that there is at least one site (Y=1), and Z=0 when there is only one site in a plot of land given that there is at least one site (Y=1). For the first stage of the analysis we had and $P(Z=0 \mid Y=1) = 1-\pi^*$. Thus, the probability of no sites located in a plot of land is $\pi \bullet P_d$, the probability of one site present is $\pi(1-\pi^*)P_d$, and the probability of more than one site present is $\pi \bullet \pi^* \bullet P_d$, where P_d is again the probability of a site being discovered.

In order to estimate π^*, we fit a logistic regression function to the data from the 29 plots where there is at least one site. A stepwise procedure again is used to select the final model. The variables x_9, number of soil boundaries, and x_{10}, distance from navigable water, are excluded from the original pool of variables since they cause great under-dispersion for this data. The details are discussed in section 4.4. Here we only note that variables are related to soil types; thus, no unique information is lost by ignoring x_9. Also, Green believes that x_{11}, the percentage of area within 1 km of navigable water, is a better indication of the importance of water routes than the absolute distance from the center of a plot, x_{10}. The model is summarized on Table 17. The significant variables are x_1 and x_{11}. Based on this model we estimate π^* as

$$\hat{\pi}^* = \left[\frac{\exp(-1.517-0.05x_1+0.043x_{11})}{1+\exp(-1.517-0.05x_1+0.043x_{11})} \right].$$

Therefore, when there is at least one site present in a plot, the data indicate that the percentage of area near navigable water greatly increases the probability of additional sites being located in the plot. This suggestive evidence supports Green's hypothesis that the proximity of navigable water influenced the probability of settlement. However, the importance of this influence is secondary to the presence of neighboring sites and fertile soil for agriculture.

Since there were only 5 plots with more than one site in the data, this second stage analysis must be considered with caution. The analysis of deviance (Table 18) shows that the model has a reasonable fit for the available data. However, in section 4.4, we show that the model has some problems with data under-dispersion. This is probably due to the lack of more observed plots with more than one site. Nevertheless, an additional advantage of this model is that it allows the estimation of a resource selection probability function to predict whether there are additional undiscovered sites in plots where only one site has been located. This estimation was not possible with the resource selection function previously proposed. The probability function is given by

$$\omega(\mathbf{x}) = \left[\frac{\exp(-1.517-0.05x_1+0.043x_{11})}{1+\exp(-1.517-0.05x_1+0.043x_{11})} \right] P_d,$$

where P_d is the probability of a site being discovered. The resource selection function

$$\omega^*(\mathbf{x})=\exp(-1.517-0.05x_1+0.043x_{11})$$

provides a selectivity index since P_d is unknown.

In our two stage analysis, then, we found evidence that agricultural considerations have the most important influence on set ement; land consisting of lime or organic soils was preferred. The presence of sites in neighboring plots also increased the probability of settlement. In the presence of at least one site in a plot, proximity of navigable water increases the probability of more sites being located in that plot.

4. Quasi-likelihood Analysis in Resource Selection Studies

In this section we introduce quasi-likelihood as a method of inference and apply it as a tool for analysis on the habitat selection studies for the spotted owl in Oregon, fernbirds in Otago, New Zealand, and the prehistoric Maya in Corozal District, Belize.

4.1 Quasi-likelihood Models

Quasi-likelihood, a method for inference developed by Wedderburn (1974), does not require that the distribution of the response variable be specified. According to McCullagh & Nelder (1983, p. 168), quasi-likelihood is equivalent to weighted least squares in the case where the weights depend only on the estimates of the regression parameters.

Suppose that a vector of responses $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ has mean $\mu=(\mu_1,\mu_2,...,\mu_n)$ and diagonal covariance matrix $\varphi V(\mu)$ whose elements are known functions of μ. For a single observation y, the individual quasi-likelihood $Q(\mu_i;y_i)$ is any solution of the differential equation

$$\frac{\partial Q(\mu_i;y_i)}{\partial \mu_i} = \frac{y_i-\mu_i}{\varphi V(\mu_i)}. \tag{19}$$

Then, the total quasi-likelihood is a sum of n components

$$Q(\mu;\mathbf{y}) = \sum_{i=1}^{n} Q(\mu_i;y_i). \tag{20}$$

For many simple variance functions, Equation 19 is easily solved. Many likelihood functions within the exponential family can be derived as quasi-likelihoods given that the appropriate variance function is assumed. Moreover, quasi-likelihood models provide the same coefficient estimates $\hat{\beta}$ as normal-theory models; however, they require only second-moment (variance) assumptions instead of the full distributional assumptions of normal-theory models (McCullagh & Nelder 1983, p. 169).

In order to obtain estimates of the coefficients $\beta'=(\beta_1,\beta_2,...,\beta_p)$, recall that within the GLM framework we defined a link function $g(\mu)$ (Equation 2) to be a known transformation of the mean μ. Because of the link function, the total quasi-likelihood (Equation 20) is a function of the regression coefficients are the values that maximize the total quasi-likelihood. These estimates are solutions to the equations

$$\sum_{i=1}^{n} \frac{y_i-\mu_i}{\varphi V(\mu_i) g'(\mu_i)} x_{ir},\ r=1,...,p,$$

where $x_{ir}=(\partial\eta_i/\partial\beta_r)$ is the value of the rth covariate.

For a response variable Y_i with mean π_i and variance $V(\pi_i)=\pi_i(1-\pi_i)$, the associated quasi-likelihood is

$$Q(\pi_i;y_i) = y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i).$$

This is the log likelihood (Equation 6) for a random variable $Y_i bin(1,\pi_i)$. In this case, however, we only assume that Y_i has variance $V(\pi_i) = \pi_i(1-\pi_i)$ instead of assuming the binomial distribution.

4.1.1 Quasi-likelihood Models for Over-dispersed Binary Data. Often the assumption of Bernoulli variation for binary data is unrealistic and over-dispersion is present. Over-dispersion occurs when variability in the data is greater than that predicted by the Bernoulli model. Again in the context of a resource selection study, let $\mathbf{y} = (y_1,y_2,...,y_n)$ be a vector of N observations for the response $\mathbf{Y} = (Y_1,Y_2,...,Y_n)$ with corresponding vector of parameters $\pi = (\pi_1,...,\pi_n)$, where Y_i, $i=1,...,n$, represents whether the ith resource unit is selected for use. Under some circumstances, we can assume

$$E(Y_i) = \pi_i \text{ and } \mathrm{Var}(Y_i) = \varphi\pi_i(1-\pi_i) \tag{21}$$

where φ is the dispersion parameter, so that $\varphi>1$ indicates over-dispersion and $\varphi<1$ represents under-dispersion relative to the Bernoulli – or Binomial(1, π_i) – model (McCullagh & Nelder 1983, p. 80). If the dispersion parameter φ is unknown, it is estimated as

$$\hat{\varphi} = 1n-p \sum_{i=1}^{n} \frac{(y_i-\hat{\pi}_i)^2}{V(\hat{\pi}_i)}.$$

Under the assumptions in Equation 21, the quasi-likelihood $Q(\pi;y)$ is equal to the Bernoulli log likelihood (Equation 7); thus, we obtain the same estimates for . Apart from the multiplier φ, quasi-likelihoods can be treated like ordinary likelihoods with the advantage that the assumptions in Equation 21 can be easily checked with graphical methods (McCullagh & Nelder 1983, p. 80).

4.2 Habitat Association Study of the Northern Spotted Owl

Due to the retrospective, observational nature of the spotted owl data, more than the expected binomial variability is likely to occur. Observational data usually have more than expected variability because unknown sources of variation cannot be controlled as in an experiment. We can incorporate data over-dispersion into the model by fitting it in terms of a quasi-likelihood function that multiplies the variance by a parameter φ known up to a constant.

Kaur et al. (1995) extended the model of Ramsey et al. as a quasi-likelihood model and conclude that the former model is inappropriate for the data. The results of the analysis are shown on Table 19. The over-dispersion parameter φ for the binomial family is estimated to be $\hat{\varphi}$=5.01. If the assumption of binomial mean to variance relation holds (Equation 4), φ is expected to be close to 1. Also, after adjusting the standard errors for over-dispersion, Wald tests on coefficients yield them insignificant; this suggests that the model is inadequate, even though the Wald test is inaccurate in the presence of correlation between parameters.

We can also adjust the modified rings model in terms of a quasi-likelihood function. The results are presented on Table . In this case, the estimated over-dispersion parameter for the binomial family is $\hat{\varphi}$=1.11, very close to the binomial model assumption of $\varphi = 1$. With regard to dispersion, then,

Table 19. Quasi-likelihood Model for Habitat Fragmentation (dispersion parameter for binomial family ($\hat{\phi}$=5.01).

Coefficients	Value	Std. Error	t value
Intercept	-0.966676872	21.656752308	-0.04463628
R_1	0.630204513	0.856140012	0.73609983
R_2	0.505918554	0.792285982	0.63855548
R_3	-0.785145053	1.080598988	-0.72658318
R_5	-0.567139585	0.580870739	-0.97636109
R_1^2	-0.008481794	0.008787443	-0.96521750
R_2^2	0.012638775	0.014308053	0.88333301
R_3^2	0.016135873	0.017957061	0.89858095
$R_1 * R_3$	0.009810176	0.011621017	0.84417533
$R_2 * R_3$	-0.036620167	0.037629360	-0.97318070
$R_3 * R_5$	0.009900098	0.009372035	1.05634456

Table 20. Modified Rings: Quasi-likelihood Model for Habitat Fragmentation (dispersion parameter for binomial family $\hat{\phi}$ =1.11).

Coefficients	Value	Std. Error	t value
Intercept	-2.672497471	6.813541384	-0.3922332
R_1^*	0.033897915	0.230840147	0.1468458
R_2^*	0.064203956	0.047336838	1.3563212
R_3^*	-0.106465795	0.281912415	-0.3776556
R_1^{*2}	-0.003313715	0.002644787	-1.2529232
R_3^{*2}	-0.004347605	0.002929868	-1.4838914
$R_1^* * R_3^*$	0.008282558	0.004378666	1.8915711

Table 21. Redefined Variables: Quasi-likelihood Model for Habitat Fragmentation (dispersion parameter for binomial family $\hat{\phi}$=1.08).

Coefficients	Value	Std. Error	t value
Intercept	-3.94264742	2.30769575	-1.7084780
x_1	0.09564732	0.13919418	0.6871503
x_2	-0.01791405	0.13673027	-0.1310174
x_3	-0.07593159	0.09989448	-0.7601180

the modified rings model seems a more adequate approach to the habitat fragmentation issue. The standard errors do not change drastically in contrast with the change in the model proposed by Ramsey et al.

Finally, when we limited our aim to explore habitat fragmentation within a 1.77 km radius around a site, we defined new variables and estimated the main effects model on Table 9. Now we also fit this model in terms of a quasi-likelihood function. The over-dispersion parameter is estimated to be $\hat{\varphi} = 1.08$. Table 21 shows that the standard errors remain the same as in the logistic regression model. Therefore, both the quasi-likelihood and the GLM approaches yield sensible analyses of habitat fragmentation in an area immediately surrounding nest sites.

4.3 Nest Selection Study of Fernbirds

The data for the fernbird nest selection study is observational. As in the spotted owl study, we may expect a dispersion parameter φ to be multiplying the binomial variance model that is assumed under logistic regression. We incorporate data dispersion into the model for nest selection by fitting it in terms of a quasi-likelihood model. The results of this analysis are shown on Table 22. The fernbird data is under-dispersed. The estimated dispersion parameter for the binomial family is $\hat{\varphi}$=0.81; we have less than the expected variability indicated in Equation 4. The cause for under-dispersion may lie in the sampling method for available sites. Harris choose 25 random sites in the study area and placed polystyrene model nests at the center of the nearest clump of

Table 22. Quasi-likelihood Model for Fernbird Nest Selection (dispersion parameter for binomial family $\hat{\varphi}$=0.81).

Coefficients	Value	Std. Error	t value
Intercept	-10.7279571	2.9776564	-3.602819
Height	7.7957204	2.9273830	2.663034
Distance	0.2097201	0.1090674	1.922849
Perimeter	0.8836828	0.4326037	2.042708

vegetation. Placing potential nests in the center of any clump may cause all non-nested sites to have very similar variable measurements (especially the distance from the nest to the outer surface of the clump) so that less than natural variation occurs. Furthermore, for this study several covariates were initially considered; hence, we can be confident that very little dispersion occurs from failure to consider significant variables that influence nest selection. It is nevertheless surprising to find under-dispersion in observational data.

4.4 Selection of Prehistoric Maya Settlements

The study on the location of prehistoric Maya sites by Green (1973) was published before the introduction by Wedderburn (1974) of quasi-likelihood inferential methods. In the previous section, we approached the issue of settlement selection using logistic regression. We could also fit a quasi-likelihood model for the data. Under this approach, the final model for selection of prehistoric Maya settlements (fit in terms of a quasi-likelihood function in Splus) is presented on Table 23. The dispersion parameter for the binomial family is estimated to be $\hat{\varphi}$=1.02; therefore, the binomial mean to variance relation from Equation 4 (assumed under logistic regression) holds. Observe that in this case, the quasi-likelihood approach yields virtually the same results (coefficient estimates and standard errors) as the logistic regression model on Table 15. We can conclude that fertile soil and presence of neighboring sites were important considerations for settlement; agricultural and social considerations probably motivated the choice of such environmental characteristics. We also see that both the GLM and the quasi-likelihood approaches are valuable for statistical analysis in this habitat selection study where a classical linear models approach is inappropriate.

We also conducted a second stage analysis to estimate the probability π^* that there is more than one Maya site in a plot of land given that there is at least one site. Through a stepwise procedure, we selected the logistic regression model presented on Table 17. The variables x_9, number of soil boundaries, and x_{10}, distance to navigable water, were excluded from the original pool of possible explanatory variables because a quasi-likelihood model including all thirteen variables yields the under-dispersion estimate $\hat{\varphi}=1.4x10^{-6}$ for the binomial family. However, when we fit a quasi-likelihood model excluding x_9 and x_{10}, the under-dispersion parameter is estimated to be $\hat{\varphi}$=0.76; this is a more reasonable initial model for a stepwise procedure.

We fit the final model from Table 17 in terms of a quasi-likelihood function; the results are shown on Table 24. The estimate of under-dispersion is $\hat{\varphi}$=0.66. The logistic regression model, then, has some problems with data under-dispersion. This under-dispersion probably occurs because the available data only includes five observations with more than one site per plot of land. The second stage analysis and the conclusions drawn from it must be considered with caution.

Given the difficulty with under-dispersion, we can suggest an alternative approach to account for the 5 plots with two Maya sites. Let the response variable Y be the count of Maya sites in a plot. The response has a Poisson distribution with intensity parameter λa, where a is the area of the plot, and the parameter λ depends upon the covariates. This

Table 23. Quasi-likelihood Model for Selection of Prehistoric Maya Settlements (dispersion parameter for binomial family $\hat{\varphi}$=1.02).

Coefficients	Value	Std. Error	t value
Intercept	-2.44627689	0.842773724	-2.902650
Lime-Soil	0.01510416	0.008720845	1.731960
Alluvial-Soil	0.01795122	0.012764965	1.406288
Marsh-Forest	-0.02838368	0.009505617	-2.985990
Neighboring Sites	0.68648031	0.222265097	3.088565

Table 24. Second Quasi-likelihood Model for Selection of Prehistoric Maya Settlements (dispersion parameter for binomial family $\hat{\varphi}$=0.66).

Coefficients	Value	Std. Error	t value
Intercept	-1.51714933	1.30075636	-1.166359
Lime-Soil	-0.05043594	0.02449963	-2.058641
Percentage near water	0.04260815	0.01710617	2.490806

method requires a reasonable model to estimate λ that is consistent with observing $Y \leq 2$ in the data. The estimated parameter $\hat{\lambda}a$ gives the expected count of sites in a plot of land. This approach is mentioned as an additional possibility for analysis, but it is not pursued in this paper.

5. Discussion

The concept of a resource selection probability function is most useful for analysis of resource selection studies. Even when the necessary sampling probabilities for estimating such a function are unknown, we can estimate a resource selection function – an index to compare resource units. This function is especially informative in studies where where the goal is to predict the future use of a resource unit. In the spotted owl study, for example, the final objective is to establish whether owls will nest in a fragmented forest habitat. Our estimated resource selection function (from the modified rings model) now provides an index to rank units based on the probability of owl nesting for different fragmentation patterns. In the study of prehistoric Maya settlements, a resource selection function is available to predict the location of undiscovered sites. Even though the data was gathered in 1973, the same approach would be ideal for analysis of updated information.

For binary response studies in particular, logistic regression, a GLM procedure available in most statistical computer packages, allows us to estimate the necessary parameters for a resource selection function. In estimating this function, however, it is crucial to identify correctly the variables that influence the probability of selection. In this paper we have used stepwise procedures, but caution is advised. Manly et al. (1993, p. 30; 1985, p. 172) and Rexstad et al. (1988) warn that stepwise regression procedures may give a biased impression of how well environmental variables explain a response phenomenon, especially when these procedures are used to chose significant variables from a larger set of potentially important explanatory variables. In the prehistoric Maya settlement study, for instance, there was an initial pool of thirteen explanatory variables, and it is possible that some important variables were excluded from the final model by the stepwise procedure. It is important to note that the selected final model is considered a good model given the available information. However, this final model is not the single best model since other approaches to model selection are available. The use of a stepwise procedure is justified by the goal of eliminating unnecessary information and reaching a parsimonious model.

Logistic regression and GLM are also useful for analysis of resource selection studies since the data are typically observational. Over-dispersion is one of the main issues to be considered in this context. Two common causes of over-dispersion are variation from relevant but ignored covariates, and unaccounted correlation between observations. These are probably some of the causes for over-dispersion in the spotted owl study; variables other than old forest density influence nest selection, and correlation between covariates occurs due to the spatial proximity of the rings. Technical issues such as wrong model selection and lack of modeling also contribute to over-dispersion.

When over-dispersion is present and GLM seem inadequate to fit the data, quasi-likelihood methods are available for analysis. In the examples presented here, quasi-likelihood provided insight into the adequacy and validity of proposed models. In the spotted owl and fernbird nesting studies, quasi-likelihood models diagnosed the presence of data over-dispersion and under-dispersion for the binomial family respectively. For the spotted owl model proposed by Ramsey et al., the quasi-likelihood approach provided adjusted standard errors that yielded many covariates insignificant and suggested the inadequacy of the model. In contrast, the fernbird model seemed reasonable as the covariates remained significant in spite of some data under-dispersion. Recall also that for the first stage location analysis of prehistoric Maya sites, virtually no data over-dispersion occured, and therefore quasi-likelihood and logistic regression yielded similar results.

In addition to the analytical approaches presented in this paper, more recent methods, developed as extensions of generalized linear models, can be applied to resource selection studies. For example, Kaur et al. (1995) have used generalized additive models (GAM) and projection pursuit regression to further analyze the spotted owl data.

References

Agresti, A. 1990. Categorical Data Analysis. John Wiley and Sons, New York.

Boyce, M. S., Meyer, J. S. & Irwin, L. 1994. Habitat-based PVA for the northern spotted owl. In: Statistics in Ecology and Environmental Monitoring, D. J. Fletcher & B. F. J. Manly (eds.) University of Otago Press, Dunedin. pp. 63-85.

Cock, M. J. W. 1978. The assessment of preference. Journal of Animal Ecology 47: 805-816.

Gionfriddo, J. P. & Krausman, P. R. 1986. Summer habitat use by mountain sheep. Journal of Wildlife Management 50: 331-336.

Green, E. L. 1973. Location analysis of prehistoric Maya sites in British Honduras. American Antiquity 38: 279-293.

Gregori, G. 1995. Ecological applications of generalized linear models and quasi-likelihood methods: an overview. Master's Paper, The Pennsylvania State University.

Harris, W. F. 1986. The Breeding Ecology of the South Island Fernbird in Otago Wetlands. Ph.D. Thesis, University of Otago, Dunedin, New Zealand.

Johnson, D. H. 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61: 65-71.

Kaur, A., Gregori, D., Patil, G. P. & Taillie, C. 1995. Ecological applications of generalized linear models and quasi-likelihood methods: an overview. Technical Report Number 95-0601, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.

Keating, K. A., and Irby, L. R. & Kasworm, W. F. 1985. Mountain sheep winter food habitats in the Upper Yellowstone Valley. Journal of Wildlife Management 49: 156-161.

Manly, B., McDonald, L. & Thomas, D. (1993). Resource Selection by Animals: Statistical design and analysis for field studies. Chapman and Hall, London.

Manly, B. F. J. 1985. The Statistics of Natural Selection on Animal Populations. Chapman and Hall, London.

McCullagh, P. & Nelder, J. A. 1983. Generalized Linear Models. Chapman and Hall, New York.

Murphy, R. K., Paine, N. F. & Anderson, R. K. 1985. White-tailed deer use of an irrigated agricultural-grassland complex in central Wisconsin. Journal of Wildlife Management 49: 125-128.

Neu, C. W., Byers, C. R. & Peek, J. M. 1974. A technique for analysis of utilization-availability data. Journal of Wildlife Management 38: 541-545.

Porter, W. F. & Church, K. E. 1987. Effects of environmental pattern on habitat preference analysis. Journal of Wildlife Management 51: 681-685.

Ramsey, F. L., McCracken, M., Crawford, J. A., Drut, M. S. & Ripple, W. J. 1994. Habitat association study of the northern spotted owl, sage grouse, and flammulated owl. Chapter 10. In Case Studies in Biometry, N. Lange, et al., (eds.), John Wiley and Sons, New York. pp. 189-209.

Rexstad, E. A., Miller, D. D., Flather, C. H., Anderson, E. M., Hupp, J. W. & Anderson, D. R. 1988. Questionable multivariate statistical inference in wildlife habitat and community studies. Journal of Wildlife Studies 52: 794-798.

Ripple, W. J., Johnson, D. H., Hershey, K. T. & Meslow, E. C. 1991. Old-growth and mature forest near spotted owl nests in Western Oregon. Journal of Wildlife Management 55: 316-318.

Ripple W. J., Lattin, P. D., Hershey, K. T., Wagner, F. F. & Meslow, E. C. 1997. Landscape composition and pattern around northern spotted owl nest sites in Southwest Oregon. Journal of Wildlife Management 61(1): 152-159.

Rolley, R. E. & Warde, W. D. 1985. Bobcat habitat use in southeastern Oklahoma. Journal of Wildlife Management 49: 913-920.

Roy, L. D. & Dorrance, M. J. 1985. Coyote movements, habitat use, and vulnerability in central Alberta. Journal of Wildlife Management 49: 307-313.

Scott, A. 1920. Food in Port Erin mackerel in 1919. Proceedings and Transactions of the Liverpool Biological Society 34: 107-111.

Stinnett, D. P. & Klebenow, D. A. 1986. Habitat use of irrigated lands by California quail in Nevada. Journal of Wildlife Management 50: 368-372.

Thomas, D. L. & Taylor, E. J. 1990. Study designs and tests for comparing resource use and availability. Journal of Wildlife Management 54: 322-330.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika 61: 439-437.