Comparation of logistic regression methods and discrete choice model in the selection of habitats

Sandra Vergara Cardozo[1]; Bryan Frederick John Manly[2]; Carlos Tadeu dos Santos Dias[3]*

[1]*Universidad Nacional de Colombia – Departamento Estadística – 111321 – Bogotá – Colombia.*
[2]*Western EcoSystems Technology Inc., Cheyenne, WY 82001 – USA.*
[3]*USP/ESALQ – Depto. de Ciências Exatas – C.P. 09 – 13418-900 – Piracicaba, SP – Brasil.*
**Corresponding author <ctsdias@esalq.usp.br>*

ABSTRACT: Based on a review of most recent data analyses on resource selection by animals as well as on recent suggestions that indicate the lack of an unified statistical theory that shows how resource selection can be detected and measured, the authors suggest that the concept of resource selection function (RSF) can be the base for the development of a theory. The revision of discrete choice models (DCM) is suggested as an approximation to estimate the RSF when the choice of animal or groups of animals involves different sets of available resource units. The definition of RSF requires that the resource which is being studied consists of discrete units. The statistical method often used to estimate the RSF is the logistic regression but DCM can also be used. The theory of DCM has been well developed for the analysis of data sets involving choices of products by humans, but it can also be applicable to the choice of habitat by animals, with some modifications. The comparison of the logistic regression with the DCM for one choice is made because the coefficient estimates of the logistic regression model include an intercept, which are not presented by the DCM. The objective of this work was to compare the estimates of the RSF obtained by applying the logistic regression and the DCM to the data set on habitat selection of the spotted owl (*Strix occidentalis*) in the north west of the United States.
Key words: resource selection, maximum likelihood, binomial distribution, comparison test

Comparação dos métodos regressão logística e modelo de escolha discreta na seleção de habitats

RESUMO: Baseado em revisão mais recente de análises de dados em seleção de recurso pelos animais e com as mais recentes sugestões, que indicam a falta de uma teoria estatística unificada que mostre como a seleção do recurso pode ser detectada e medida, os autores sugerem que o conceito da função da seleção do recurso (RSF) pode ser a base do desenvolvimento da teoria. A revisão de modelos de escolha discreta (DCM) é sugerida como uma aproximação para estimar a RSF quando a escolha do animal os grupos de animais envolvem diferentes conjuntos de unidades de recurso disponíveis. A definição do RSF requer que o recurso que esteja sendo estudado consista em unidades discretas. O método estatístico frequentemente usado para estimar a RSF é a regressão logística mas DCM também pode ser usado. A teoria de DCM tem sido bem desenvolvida para análises de conjunto de dados que envolvem escolhas de produtos pelos humanos, mas também pode ser aplicável a escolhas de habitat pelos animais com algumas modificações. A comparação da regressão logística com o DCM para uma escolha é feita porque as estimativas do coeficiente do modelo de regressão logística inclui o intercepto, mas no DCM o coeficiente do intercepto não está presente. O objetivo deste trabalho foi comparar as estimativas da função da seleção do recurso obtida pela aplicação da regressão logística e o DCM do conjunto de dados de um estudo de seleção de habitat da coruja manchada (Strix occidentalis) no noroeste dos Estados Unidos.
Palavra-chave: seleção de recurso, máxima verossilmilhança, distribuição binomial, testes de comparação

Introduction

Natural resources include materials found in nature that permit a species to survive. These resources can be renewable or non renewable. Animal populations need these resources to survive. Differential selection of available resources is one of the primary factors that allow species to co-exist, and is therefore a priority in the preservation of endangered species (Rosenzweig, 1981). Consequently, an adequate supply of natural resources is needed to sustain animal populations, and when a species better selects its resources, the better.

Under certain assumptions, the population density of an animal species depends on the availability of a resource in equilibrium (Fagen, 1988). Resource selection (RS) is used in studies to identify resources critical to an animal population and to predict the incidence of the species. Frequently, animals are monitored individually and then grouped to estimate effects at a population level. Resource selection functions (RSFs) are statistical models that require the variables under study to consist of discrete units. The theory of discrete choice models has been well developed for analyses of human choice data (Train, 2003). McDonald et al. (2006) suggest that

these models may also be modified to be applied to animal choices. These authors were motivated by studying the comparison of the discrete choice model with the logistic regression model and in this way compare the coefficient estimates. Here the RSF is compared with the exponential resource selection function (RSF).

Material and Methods

This study utilizes data from nocturnal activities of the spotted owl (*Strix occidentalis*), collected in two discrete areas (Klamath and Korbel) within the property of the Green Diamond Resource Company (GDRCo) in Del Norte and Humboldt countries of northwestern California, USA.

Twenty-eight areas occupied or used by owls during the nocturnal period were identified between April 1998 and September 2000, using radio-telemetry. McDonald et al. (2006) used back-pack harness mounted radio-transmitters, and in this way it was possible to verify that five owls resided in Klamath and twenty-three in Korbel. Forty-six explanatory variables were simultaneously observed (Table 1), resulting in a total of 8,739 observations (Ryan, 2004; McDonald et al., 2006).

According to McDonald et al. (2006), applications of discrete choice models generally assume that animals make a series of choices based on a finite set of discrete habitat units, known as choice sets. Other resource selection analyses include logistic regression that is applied to a sample of used and not used resource units and assumes that choices are made from a set of available resource units.

Discrete choice models (DCM) are usually applied in situations where n sets of resource units, n_i (i = 1,

Table 1 - Explanatory variables for the spotted owl (*Strix occidentalis*).

Variable	Description
aclc2	indicator variable for age class of stand = 6 to 20 years
aclc3	indicator variable for age class of stand = 21 to 40 years
aclc4	indicator variable for age class of stand = 41+ years
acln2	indicator variable for age class of nearest stand = 6 to 20 years
acln3	indicator variable for age class of nearest stand = 21 to 40 years
acln4	indicator variable for age class of nearest stand = 41+ years
(*aclc_*)x(*acln*)	9 indicator variables for the interaction of (age class of stand) by (age class of nearest stand)
acl_p2	Percentage of buffer aged 6+ to 20 years
acl_p3	Percentage of buffer aged 21+ to 40 years
acl_p4	Percentage of buffer aged 41+ years
age_cent	age stand of trees (years)
acl_dne	distance to nearest edge (ft) [1]
acl_ed	Edge density (ft/acre) in circular buffer (edge defined as change in age class)
acl_mps	Mean patch size (acre) in circular buffer (patch defined as uniform age class)
acl_np	number of patches in circular buffer
acl_pd	patch density (n/100 acres) in circular buffer
acl_pscv	patch size CV (%)
acl_te	Total edge in circular buffer
drd_cent	distance to mainline road (ft)
hgt_cent	height of trees (ft)
phw_cent	% hardwood
prs_cent	% residual
prw_cent	% redwood
hwd_cent	hardwood basal area (ft^2/acre)
rwd_cent	redwood basal area (ft^2/acre)
wwb_cent	whitewood basal area (ft^2/acre)
tba_cent	Total basal area (ft^2/acre)
specent2	indicator variable for dominant species = redwood
specent3	indicator variable for dominant species = whitewood
slp_cent	slope position (%)

Source: Western EcoSystems Technology, Inc - Cheyenne WY – USA. [1] ft = 30.48 cm. [2] acre = 4047 m^2

2,..., n), are defined as available for selection, and one unit is chosen from each of the choice sets. It is assumed that the probability of selecting the j^{th} unit of the i^{th} choice set is proportional to $\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp})$, in which $\beta_1, \ldots \beta_p$ are coefficients to be estimated and $x_{ij1}, \ldots, x_{ijp}$ are values of *p* covariates measured in the j^{th} unit of the i^{th} choice set. The probability of the j^{th} unit being selected from the i^{th} choice set is then:

$$p_{ij} = \frac{\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp})}{\sum_{k=1}^{n_i} \exp(\beta_1 x_{ij1} + \ldots + \beta_p x_{ikp})}. \qquad (1)$$

Then for *S* independent choices, the likelihood function is equal to the product of the probabilities of the successful choices.

$$L = l(\beta_1, \beta_2, \ldots, \beta_p) = \prod_{i=1}^{S} (p_{i1})^{y_{i1}} \times (p_{i2})^{y_{i2}} \times \ldots \times (p_{in_i})^{y_{in_i}} \qquad (2)$$

in which $y_{ij} = 1$, if the *j*th resource unit is chosen for choice set *i* and $y_{ij} = 0$, otherwise n_i is the choice unit of the i^{th} choice set, and p_{ij} is the value given by the expression (1).

Maximum likelihood estimates of the parameters β are obtained by maximizing **L** with respect to these parameters. This also, provides estimates of standard errors and allows significance tests.

According to Manly et al. (2002), In this case logistic regression can be used to relate the probability of use of variables x_1 to x_p that are measured on the resource units.

The logistic regression is a special case of DCM that allows for a binary choice. The RSPF resource selection probability function is simply assumed to take the form,

$$w^*(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \qquad (3)$$

In this case of DCM with one choice unit (available or used), the probability of using the resource is

$$p_1(x) = \frac{\exp(\beta_0 x_{10} + \beta_1 x_{11} + \ldots + \beta_p x_{1p})}{\exp(\beta_0 x_{00} + \beta_1 x_{01} + \ldots + \beta_p x_{0p}) + \exp(\beta_0 x_{10} + \beta_1 x_{11} + \ldots + \beta_p x_{1p})}$$

This probability can be rewritten as,

$$p_1(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}.$$

by letting $x_i = x_{1i} - x_{0i}$, i = 0,...,p, where $x_{10} = 1$ and $x_{00} = 0$

in which $x = (x_1, \ldots, x_p)$ is the vector of values of the explanatory variables X. The logistic function has the desired property of restricting the probability values of w*(x) to between 0 and 1.

When using logistic regression with census data the assumption made is that there are N available resource units and it is known which of these have been used and which have not been used after a single period of selection, Manly et al. (2002).

Another justification for using the logistic function rather than other approaches to approximate RSPF is the fact that it is widely used for other statistical analyses in biology; consequently, several computer programs are currently available to estimate these parameters.

The estimated function, $\hat{w}(x) = \exp(\hat{\beta}_1 x_{ij1} + \hat{\beta}_2 x_{ij2} + \ldots + \hat{\beta}_p x_{ijp})$ is then the RSF, gives the relative probability of use of different types of resource units. Computer software packages that estimate discrete choice model parameters by maximum likelihood include SAS/Proc PHREG and S-Plus routine COXPH, (Manly et al., 2002).

When a parametric model for RS probability is used, parameters are estimated by the maximum likelihood. Therefore the quantity,

$$D = -2\{\log_e(L_p)\} + 2p, \qquad (4)$$

is called the deviance, which can be used as a measure of the agreement of the model, *p* is the number of unknown parameters in the model to be estimated (Akaike, 1974). If L_M is the maximum likelihood of the adjusted model, and L_F ($\geq L_p$) is the likelihood of the model perfectly fitting the data, then $L_F = L_p$ corresponds to a Null Model (N.M).

Chi-squared tests of deviance may be used to evaluate the evidence of the probability of use in the study areas. Under certain distribution conditions, deviance statistics approximately follow a Chi-squared distribution with the degrees of freedom (df) defined by the number of observations less the number of parameters estimated. Deviance is analogous to the sum of squares in regression models of analysis of variance.

Design II (Manly et al., 2002) was used on the spotted owl data, in which animals are identified individually and use of the resource units is measured for each individual, but availability of the resource is measured for the whole population. Sample protocol C was used in which the resource units used and not used are sampled independently (Manly et al., 2002).

Logistic regression can still be used in this case. However, a special justification is needed depending on the types of samples involved. In the present case, for independent samples of used and available units, a population of available units of size N is assumed, with the i^{th} unit assuming values $x_i = (x_{i1}, \ldots, x_{ip})$ for variables X_1 to X_p and the relative probability of use of the different resource units corresponding to:

$$w^*(x) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \qquad (5)$$

The sampling plan is such that each available unit has a probability P_a of being sampled, and each used unit has a probability P_u of being sampled, with a sample of available units selected first with no replacement so that the units in this sample cannot appear in the sample of not used units. In this case the probability of a unit being used and sampled is $(1 - P_a)w^*(x_i)P_u$ and the probability of a unit being in the sample of used units or in the sample of available units is given by:

$$\text{Prob(ith unit sample)} = P_a + (1 - P_a)w^*(x_i)P_u \quad (6)$$

Consequently, the probability of the i^{th} unit being in the sample of used units, given that it was sampled is given by:

Prob(ith unit used/ sampled) = Prob (used and sample) / Prob (sampled)

$$= \frac{(1-P_a)w^*(x_i)P_u}{P_a + (1-P_a)w^*(x_i)P_u} \quad (7)$$

Given that the RSPF defined in equation (5) assumes a particular exponential form, the probability of expression (7) may also be written as:

$$\tau(x_i) = \frac{\exp\left\{\log\left[\frac{(1-P_a)P_u}{P_a}\right] + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right\}}{1 + \exp\left\{\log\left[\frac{(1-P_a)P_u}{P_a}\right] + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right\}} \quad (8)$$

This corresponds to an expression of logistic regression in which the parameter β_0 is modified as follows, to allow for the sampling probabilities of available and used units:

$$\beta'_0 = \log\left[\frac{(1-P_a)P_u}{P_a}\right] + \beta_0 \quad (9)$$

Assuming independent observations, $\mathbf{x}_i$ represents the probability of observing resource unit *i* as being used, and the probability of observing that same unit as being available given by $1 - \tau(x_i)$. Let y_i be the indicator of use or non-use of a sample unit, so that $y_i = 0$ for sampled unit *i* pertaining to the sample of available units and $y_i = 1$ for sampled unit *i* pertaining to the sample of used units.

The probability of observing unit *i* could then be written as,

$$L_i = \tau(x_i)^{yi}\,(1 - \tau(x_i))^{1-yi} \quad (10)$$

and the logarithm of the likelihood of observing the complete sample is:

$$\log\{L(\beta_0, \beta_1, \ldots, \beta_p)\} = \sum_{i=1}^{n} \log L_i \quad (11)$$

$$\log\{L(\beta_0, \beta_1, \ldots, \beta_p)\} = \sum_{i=1}^{n} y_i \log\{\tau(x_i)\} + (1 - y_i)\log\{1 - \tau(x_i)\} \quad (12)$$

Computer programs for logistic regression can be used to estimate coefficients β_0, β_1, ..., β_p of the linear logarithm function of the expression (5).

The fact that the logistic regression constant β'_0 assumes the expression form (9) means that if the probabilities of samples P_u and P_a are known, then the parameter b_0 of RSPF in the expression (5) can be estimated subtracting the quantity $\log\left[\frac{(1-P_a)P_u}{P_a}\right]$ from the constant estimated in the logistic regression equation. If the sampling fractions are not known, then b_0 cannot be estimated; however, it is still possible to estimate RSF,

$$w^*(x) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p) \quad (13)$$

and use this function to compare resource units.

Note that the correct relative probabilities of use are obtained by substituting estimates of β_0, β_1, ..., β_p in the linear logarithm function of expressions (5) or (13). The probabilities obtained using computer programs to adjust the logistic regression $\tau(x_1)$ in expression (8) are not correct estimates for selecting the probability of resource $w^*(x_i)$, or for the resource selection function $w(x_i)$, since the total number of units used by the animals is not assumed to be known.

Results and Discussion

To compare the logistic regression and the discrete choice model, a random sample of 390 observations was selected from the spotted owl data with one choice. Variable selection followed the Akaike information criteria (AIC).

Minitab (1997) and The R Development Core Team (2006) software's were used for the logistic regression estimates, and Fortran programming language (Fortran, 1977) was used to estimate the DCM parameters.

The adjustment of the binomial distribution with a logit link function for the selected variables can be seen in Figure 1. The "worm" graph in Figure 2 is a general diagnostic tool for residual analyses. The vertical axis represents the differences between theoretical and empirical distributions. The "worm" graph should be in the form of a cord, indicating a binomial data distribution in the present case in which consecutive points can be observed (Buuren and Fredricks, 2001). Figures 2 show that the worm graph of the binomial distribution with a logit link function is not adjusted very well.

The parameter estimates for the logistic regression and discrete choice model are shown in (Table 2).

The comparison of the estimates of the two methods differs with respect to the intercept. However, when analyzing data of the behavior of the animals it is a little difficult to interpret the intercept in the logistic regression model. Here the discrete choice model is proposed with one choice for the analyses of data from animals. With a discrete choice model for resource selection, the *i-th* choice is described by the choice set of resource units (habitat or food) that are available to be chosen; and values for variables that characterize all resource units in the choice set (e.g., vegetation type, elevation, etc.).

A comparison of models by the chi-squared test using the logistic regression and discrete choice model parameters is shown in (Table 3). This table shows the deviance of DCM is -56.76 less than that of RL, and AIC of DCM is -58.76 less than that of RL.

Note that logistic regression has one degree of freedom less than DCM since the latter does not have an intercept.

In Figure 3, we can see that the spotted owl visited many places, although it used few of them. The situation is more complex for the independent random

Table 2 – Estimates of Logistic Regression and Discrete Choice Model parameters (DCM) for habitat selection of the spotted owl.

	Logistic Regression			Discrete Choice Model	
Variables	Coefficient	Std. err[1]	P-value[2]	Coefficients	Std. err.
Constant	-3.9758	1.2618	0.002	-	-
aclc2	-0.3606	1.1461	0.753	-0.0747	1.2672
aclc3	-1.0449	1.4836	0.481	-1.1875	1.5949
aclc4	0.4724	1.1233	0.674	0.7734	1.2807
acl_ed	0.0127	0.0058	0.029	0.0176	0.0077
hgt_cent	0.0104	0.0049	0.034	0.0147	0.0064
slp_cent	-0.0171	0.0073	0.019	-0.0289	0.0103

[1]Estimated standard errors output from the fitting process. [2]The p-values shown are obtained by calculating the ratios of estimates to their standard errors and finding.

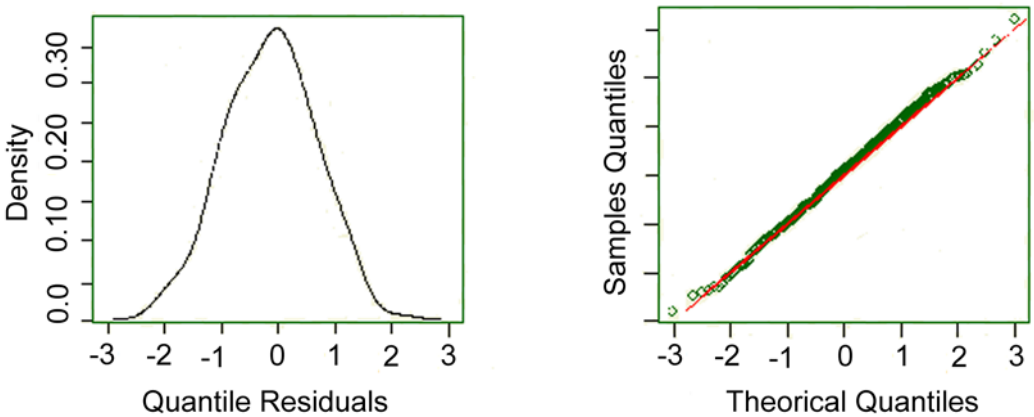

Figure 1 – Distribution of residual frequencies and QQ plot for Binomial distribution with logit link function.

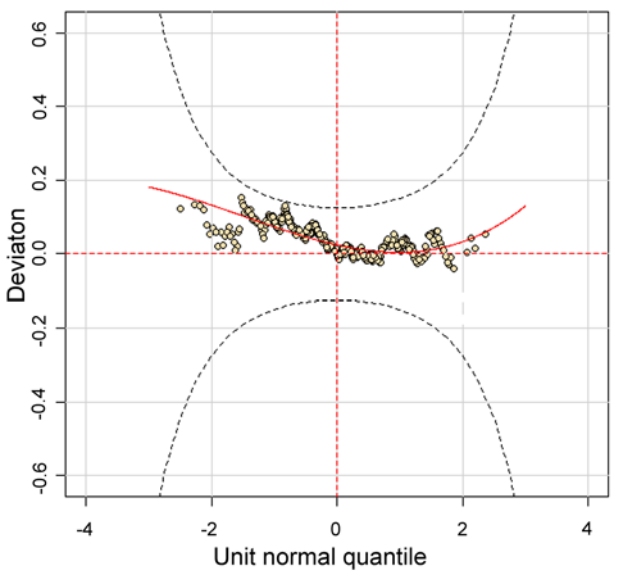

Figure 2 – Worm graph of binomial distribution with logit link function.

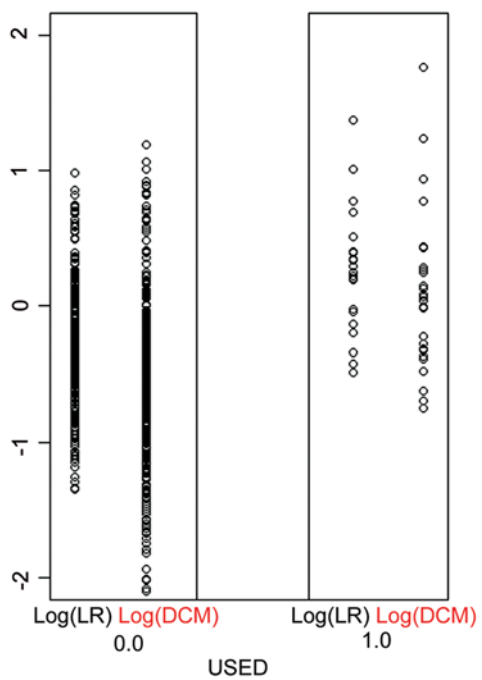

Figure 3 – Comparison of uses of the Spotted Owl (*Strix occidentalis*) with logistic regression and discrete choice model (DCM).

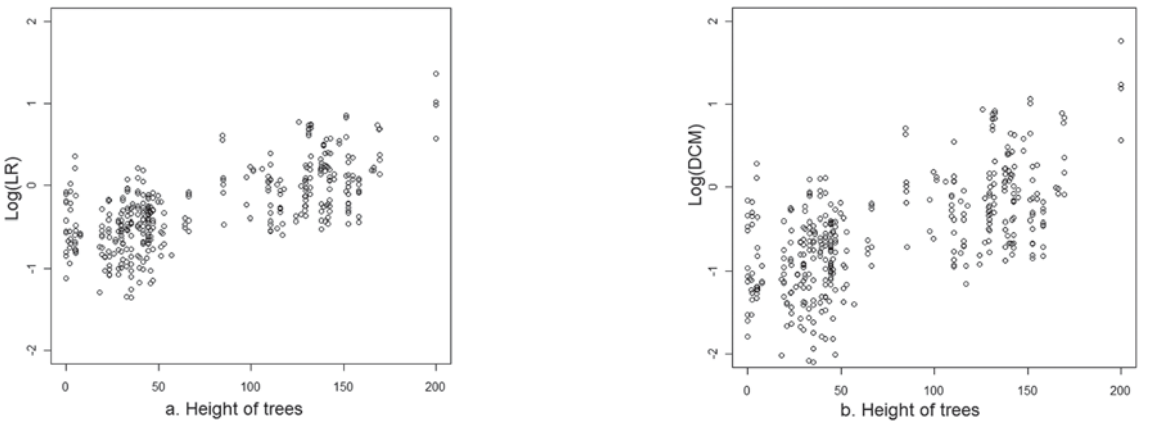

Figure 4 – a. Height of trees vs Log (Logistic Regression), b. Height of trees vs Log (Discrete Choice Model (DCM).

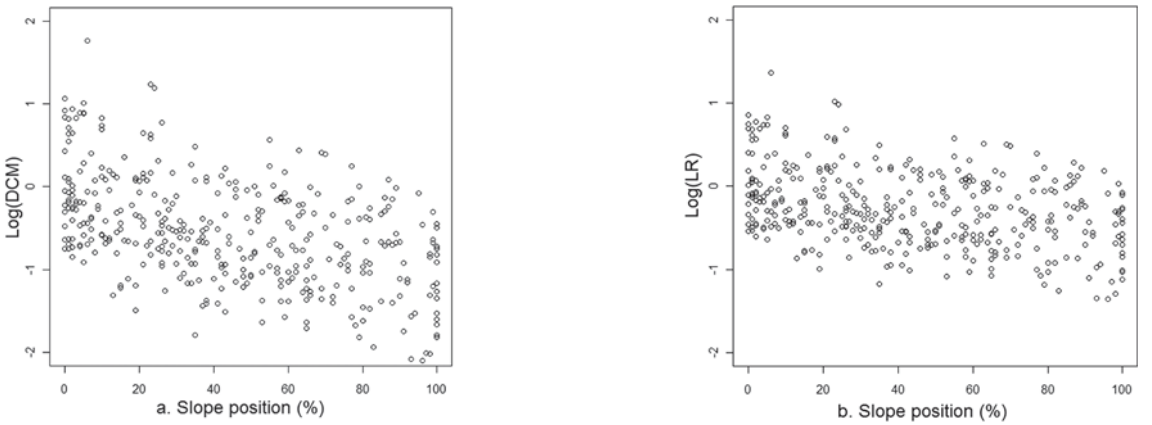

Figure 5 – a. Slope position vs Log (Discrete Choice Model (DCM), b. Slope position vs Log (Logistic Regression).

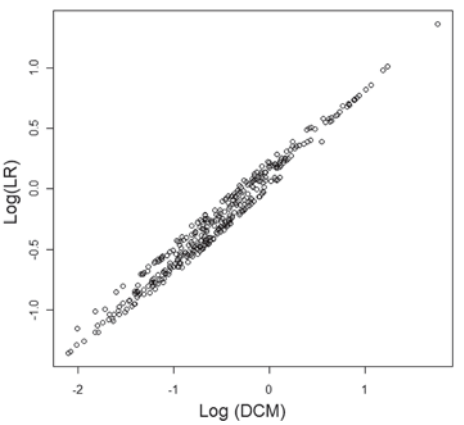

Figure 6 – Log (Discrete choice model (DCM)) vs Log (Logistic Regression).

Table 3 – Comparison of models by the Q-squared test.

	Logistic Regression		Discrete Choice Model (DCM)	
Deviance	163.78	(383 df)	107.02	(384 df)
AIC	177.78		119.02	

samples that were taken separately of different unit types: available, used and non-used.

The similarity between the logistic regression and DCM graphs shown in (Figure 4) should be observed, particularly the variable height of trees vs log (logistic regression) and the height of trees vs log (DCM) a light dispersion in the graph of the logistic regression.

In a same way it should be observed in Figure 5 (a) that the graph of the variable slope position vs log (Logistic regression) and in Figure 5 (b) slope position vs log (DCM) a slight dispersion in the graph of the logistical regression.

In estimating each of the owl choices, it can be observed that the coefficients estimated for the logistic regression differ from the coefficients of all owl choices. The same occurred with DCM in estimates of choice model parameters for all owls. In the graph of the estimates of logistic regression and DCM there is a better adjustment with respect to DCM, and Table 3 indicates the best estimate of the deviance and Akaike information criteria (AIC) of the DCM model (Figure 6).

Conclusions

• Resource selection functions estimated by logistic regression successfully identified the resources critical to an animal population and predicted the occurrence of species in different locations.

• An adjusted logistic regression and the discrete choice

model (DCM) were the best methods for predicting choices of the spotted owl.

• Parameter estimates for logistic regression and DCM with a one choice hadsimilar performances.

• An analysis made of all choices together differed from the analyses made choice by choice, justifying the use of random effect models for all animals considered simultaneously. However, logistic regression and DCM can be generalized to include random effects.

References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

Buuren, S.V; Fredricks, M. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. Statistics in Medicine, 20: 1259-1277.

Fagen, R. 1988. Population effects of habitat change: a quantitative assessment. Journal of Wildlife Management 52: 41-46.

FORTRAN 77. 1995. Programmer's Guide. Wadsworth Pub. Co., Belmont, CA, USA.

Manski, C. 1988. Structural models for discrete data: the analysis of discrete choice. p.58-109. In: Leinhardt.S., ed. Sociological methodology. Jossey-Bass, San Francisco, CA,USA.

McCracken, M.L.; Manly, B.F.J.; Vander-Heyden, M. 1998. The use of discrete: choice models for evaluating resource selection. Journal of Agricultural, Biological, and Environmental Statistics 3: 268-279.

Manly, B.F.J.; McDdonald, L.L.; Thomas, D.L.; McDdonald, T.L.; Erickson, W.P. 2002. Resource Selection by Animals. 2ed. Kluwer Academic, London, UK.

McDonald, T.L.; Manly, B.F.J.; Nielson, R.M.; Diller, L.V. 2006. Discrete-choice modeling in wildlife studies exemplified by Northern Spotted Owl nighttime habitat
selection. Journal of Wildlife Management 70: 375-83.

MINITAB. 1997. Minitab User's Guide 2: Data Analysis and Quality Tools. Minitab Inc., State College, PA, USA.

The R Development Core Team, 2006. R: A Language and Enviroment for Statistical Computing, Vienna, Austria. Available in http://www.R-project.org. [Accessed May 01, 2006].

Rosenzweig, M.L. 1981. A theory of habitat selection. Ecology 62: 327-335.

Ryan, N.; McDonald, T.L; Lamphear, D. 2004. Northern Spotted Owl Nighttime Site Selection Model. Report Western EcoSystems Technology,Cheyenne, WY, USA.

Train, K. 2003. Discrete Choice Methods with Simulation. University Press,Cambridge, UK.

Received December 19, 2008
Accepted December 01, 2009