ESTIMATION OF RESOURCE SELECTION FUNCTIONS WHEN USED AND AVAILABLE SAMPLES OVERLAP

TRENT L. McDONALD[1], Western EcoSystems Technology, Inc., 2003 Central Ave., Cheyenne, WY 82001, USA

Abstract: The most popular sampling scheme for the study of resource selection calls for data to be collected from two separate samples; one containing used resource units, the other containing available resource units. If one or more resource units appear in both the used and available sample, the samples are said to overlap, and questions have arisen regarding how to properly estimate the resource selection function in this case. One approach assumes that the available sample was chosen first, without replacement, and therefore units that happen to be selected by both samples are considered to be members of the available sample only. This approach causes difficulty in some situations because researchers typically have an abundance of available units, but few used units. The approach taken here assumes that used points were sampled first, without replacement, and therefore overlap units belong in the used sample only. This paper derives the statistical likelihood for the latter approach and shows that coefficients can be estimated using one of three analyses. The particular analysis approach taken depends upon whether or not sampling probabilities for used and available units are known, and assumptions regarding the size of these sampling probabilities.

Key words: RSF, habitat, mapping, GIS

Resource selection studies (Manly et al., 2002) compare characteristics of used or consumed resource units (e.g., plots of land, denning or nesting locations, prey or food items, etc.) with characteristics of resource units that have not been used. This comparison of used and not used resource units is fundamentally based on the frequencies of characteristics in each group, and generally results in statements or predictions of the relative probability of a resource unit being used given its characteristics. Such statements are useful for a variety of reasons, including prediction of future locations, defining high and low quality habitat, quantitative risk assessment, assessing relative trade-offs of changes in one characteristic over another, etc.

In most resource selection studies, the population of resource units is thought of as being partitioned into two types; those used and those unused. In some research situations, it is possible to draw a single random sample of resource units and determine which have been used and which have not. In this case comparison of used and unused unit characteristics is straightforward using logistic regression. However, more often a sample of used resource units is collected separately from a sample of unused or available units because it is generally cost prohibitive to collect a large enough sample of resource units to insure that both types of units are well represented. It is generally easier to collect a sample of used units by following an animal or population through time, and then another sample of available (or random) units at a separate time. Studies where animal locations are collected by radio-telemetry or aerial survey techniques often fall into this category because the sample of available locations is collected by a separate survey or from a map in a geographic information system (GIS).

In studies where one sample of used and one sample of available units are taken, it is possible for a single resource unit to appear in both the used and available samples. When one or more units appear in both the used and available samples, the samples are said to overlap, and the unit(s) included in both are called overlap units. It is often the case that the probability of overlap is near zero, particularly when data are derived from a GIS; nonetheless, in practice questions have arisen regarding what to do with overlap units. The primary questions that arise are whether overlap units should be included in the used sample only, the available sample only, or both, and what is the proper analysis in each case.

The approach to dealing with overlap advocated by Manly et al. (2002, section 5.4) considers overlap units to be members of the available sample only. The Manly et al. approach is justified by assuming that the available sample was taken first, without replacement, and therefore any unit in the available sample cannot also appear in the used sample. Under this assumption, logistic regression can be used to estimate coefficients in a resource selection function (RSF, Manly et al. 2002) of the form $\exp(b_1x_1 + \ldots + b_px_p)$ (because probabilities are relative, an intercept is generally not included in the resource selection function). However, this approach treats potentially expensive-to-collect used units the same as inexpensive and potentially abundant available units. Typically, many available units but few used units are collected, and researchers justifiably want to treat all used units as used units. As a result, this approach, while valid, is unpalatable in some situations.

The approach taken by this paper is the reverse of the approach of Manly et al. (2002). Here, it is assumed that the used sample was taken first, without replacement, and any unit in the used sample cannot also appear in the

[1] Email: tmcdonald@west-inc.com

available sample. Under this assumption, overlap units are members of the used sample only. This approach alleviates uneasiness associated with excluding certain used units from the used sample simply because they happen to be included in the sample of available units. If a unit selected into the available sample subsequently turns out to be used, and if it does not also appear in the sample of used units, the unit is not in the overlap and should be kept in the sample of available units.

In the next section, a statistical likelihood for the situation where overlap units are included in the sample of used units is derived and three potential analyses are described. Which analysis to use in a particular situation depends on which sampling probabilities are known or assumed to be near 0. If both the probability of sampling a used unit and the probability of sampling an available unit are known, commonly available generalized linear model software can be used to maximize the likelihood. If only the probability of sampling an available unit is known, logistic regression with an offset term can be used to maximize the likelihood in most situations assuming the RSF has a particular form. If sampling probabilities are unknown, but the probability of sampling an available unit is small, logistic regression can again be used to maximize the likelihood in most situations assuming a different functional form for the RSF. A discussion of the assumptions and conditions necessary for each analysis to be valid is included in the last section.

METHODS

Assume that used units were sampled without replacement prior to drawing the sample of available units. If a unit in the used sample happens to also appear in the sample of available units, the unit should be considered a member of the used sample only. If it becomes known that a unit appearing in the sample of available units only was in fact used, it should remain in the available sample because theory behind the likelihood requires the possibility of obtaining used units in the available sample. If it were impossible to obtain a used unit in the available sample, the available sample would be a sample of unused units.

Let the vector of characteristics associated with unit i be $\mathbf{x}_i = [x_{i1}\ x_{i2} \ldots x_{ip}]$. Let $w^*(\mathbf{x}_i b) = w^*(\mathbf{b}_0 + x_{i1}\mathbf{b}_1 + x_{i2}\mathbf{b}_2 + \ldots + x_{ip}\mathbf{b}_p)$, where b is a vector of coefficients, represent a resource selection probability function (Manly et al., 2002). Then, by definition, $w^*(\mathbf{x}_i b)$ represents the probability that unit i with characteristics $\mathbf{x}_i$ is used. Let the probability of including a particular used unit in the used sample be P_u. Under these assumptions, the probability of unit i being used and included in the sample of used units is $w^*(\mathbf{x}_i b)P_u$.

The probability of a unit appearing in the sample of available units is the probability of it not appearing in the sample of used units and subsequently appearing in the sample of available units. The probability of including a particular unit in the sample of available units is $(1 - w^*(\mathbf{x}_i b)P_u)P_a$, where P_a is the probability that a unit not drawn by the sample of used units is drawn by the sample of available units.

The statistical likelihood of data in the combined used and available samples is derived by conditioning on the samples and computing the probability of a unit being used given that it was included in one of the two samples. Since a single unit cannot appear in both samples, the probability of a unit appearing in the combination (union) of the two samples is $(1 - w^*(\mathbf{x}_i b)P_u)P_a + w^*(\mathbf{x}_i b)P_u$. The probability of a unit being used given that it was sampled is,

$$
\begin{aligned}
P(unit\ i\ used|sampled) &= \frac{w^*(\mathbf{x}_i b\ )P_u}{(1\text{-}w^*(\mathbf{x}_i b\ )P_u\ )P_a + w^*(\mathbf{x}_i b\ )P_u} \\
&= \frac{w^*(\mathbf{x}_i b\ )P_u}{P_a + (1\text{-}P_a\ )w^*(\mathbf{x}_i b\ )P_u}.
\end{aligned} \tag{1}
$$

Now let the random variable y_i equal 1 if unit i in the combined sample came from the sample of used units, and 0 if unit i was a member of the sample of available units. If $t_i = P(unit\ i\ used|sampled)$, the full likelihood of the data is,

$$
L(\ b\ ) = \prod_{i=1}^{n} t_i^{\,y_i}(1\text{-}t_i)^{1\text{-}y_i},
$$

where n is the number of units in the combined sample. Taking logarithms,

$$\ln(L(b)) = \sum_{i=1}^{n} y_i \ln(t_i) + (1\text{-}y_i)\ln(1\text{-}t_i)$$

$$= \sum_{i=1}^{n} y_i \ln\left(\frac{t_i}{1\text{-}t_i}\right) + \ln(1\text{-}t_i).$$

Regardless of the form of w^*, this likelihood is identical to the likelihood of n Bernoulli trials; i.e. it is the binomial likelihood with the number of trials set to 1. If P_u and P_a are known and the link function (McCullagh and Nelder, 1989, p. 27) is defined correctly, maximum likelihood estimates of b can be obtained from a generalized (binomial) linear model. Use of a generalized linear model when P_u and P_a are known allows estimation of the absolute probability of selection for all units for which the covariates are known. In addition, use of a generalized linear model allows w^* to take on any functional form. Useful functional forms for w^* usually guarantee $0 < w^* < 1$, which is not necessarily the case when logistic regression is used (see other two analyses below). To obtain maximum likelihood estimates using a generalized linear mo del, the link function should be defined as

$$g(t) = w^{*-1}\left(\frac{tP_a}{P_u[1\text{-}t + tP_a]}\right) = w^{*-1}\left(\frac{t}{[1\text{-}(1\text{-}P_a)t]} \cdot \frac{P_a}{P_u}\right), \tag{2}$$

where w^{*-1} is the functional inverse of w^*. The link $g(t)$ was obtained by substituting t into Equation (1) and solving for $\boldsymbol{x}_i b$. Link functions are required to be monotonic and differentiable (McCullagh and Nelder, 1989, p. 27), and provided w^* is monotonic and differentiable, $g(x)$ is also monotonic and differentiable.

Many generalized linear model software packages, such as SAS Proc Genmod and the S-Plus function glm, allow users to define custom link function and therefore can be used to obtain maximum likelihood estimates of b when P_u and P_a are known. Once estimates of b are obtained, they can be substituted into $w^*(\boldsymbol{x}_i b)$ to obtain estimates of the probability of use for each unit in the population. If w^* is not monotonic or differentiable, maximum likelihood estimates of b can still be obtained by directly maximizing the likelihood using an algorithm such as Newton-Raphson.

The above method for estimating b requires that both sampling probabilities (P_u and P_a) be known. This requirement significantly impairs application of the method because sampling probabilities are rarely known in practice. The remainder of this section relaxes this requirement to accommodate the following two cases: (1) P_a known but P_u unknown, and (2) P_a and P_u both unknown and P_a is near 0. In both cases, an explicit functional form for w^* is assumed, and only relative, rather than absolute, probabilities can be estimated. Both cases require a constraint on $\boldsymbol{x}_i b$ that limits application of the method; however, these constraints are usually satisfied.

When P_a is known but P_u is not known, estimation of b can be accomplished by assuming a specific form for w^* in which P_u cancels, and then recognizing that the resulting function is proportional to w^*. Assuming temporarily that both P_a and P_u are known let w^* be the scaled logistic function,

$$w^*(\boldsymbol{x}_i b) = \frac{exp(\boldsymbol{x}_i b)}{P_u[1 + exp(\boldsymbol{x}_i b)]} \tag{3}$$

where $\boldsymbol{x}_i b < \log(P_u/(1\text{-}P_u))$. The reason for this constraint on $\boldsymbol{x}_i b$, and how it limits application of the method, is discussed in the next section. Substituting this form of $w^*(\boldsymbol{x}_i b)$ into Equation (1) yields,

$$t_i = P(unit\ i\ used|sampled) = \frac{\dfrac{exp(\boldsymbol{x}_i b)}{P_u[1 + exp(\boldsymbol{x}_i b)]} P_u}{P_a + (1\text{-}P_a)\dfrac{exp(\boldsymbol{x}_i b)}{P_u[1 + exp(\boldsymbol{x}_i b)]} P_u}$$

$$= \frac{exp(\boldsymbol{x}_i b)}{P_a[1 + exp(\boldsymbol{x}_i b)] + exp(\boldsymbol{x}_i b)[1 - P_a]}$$

$$= \frac{exp(\boldsymbol{x}_i b)}{P_a + exp(\boldsymbol{x}_i b)}$$

$$= \frac{exp(-\ln(P_a) + \boldsymbol{x}_i b)}{1 + exp(-\ln(P_a) + \boldsymbol{x}_i b)}.$$

Due to the form of t_i in this case, and the binomial nature of the likelihood, estimates of b can be obtained using logistic regression. Note that P_u does not appear in the expression for t_i and does not influence estimation of the coefficients. The logistic regression used to estimate b should contain the offset term $-\ln(P_a)$ (McCullagh and Nelder, 1989, p. 206) in order to properly estimate the intercept term b_0, and should assume the standard logistic link function. If P_a were not known, the logistic regression routine's intercept parameter would estimate the quantity $[-\ln(P_a) + b_0]$, and it would not be possible to estimate b_0 separately from $-\ln(P_a)$. If b_0 cannot be properly estimated, neither w^* nor a function proportional to w^* can be estimated due to non-proportionality inherent in the assumed logistic form for w^*. Nonetheless, when P_a is known and used as an offset it is possible to estimate a function proportional to w^* (i.e., the RSF) by substituting the estimated coefficients into Equation (3), and dropping P_u.

When P_a and P_u are both unknown, but P_a is near 0, estimation of b can again be accomplished by assuming a specific form for w^* and recognizing that logistic regression estimates a function proportional to w^*. Let w^* have the exponential form,

$$w^*(\boldsymbol{x}_i b) = exp(\boldsymbol{x}_i b) = exp(ß_0)exp(x_{i1}ß_1 + \ldots + x_{ip}ß_p)\,, \tag{4}$$

where $\boldsymbol{x}_i b < 0$. Again, the reason for this constraint on $\boldsymbol{x}_i b$ and how it limits application of the method is discussed in the next section. Substituting this form for $w^*(\boldsymbol{x}_i b)$ into Equation (1) yields,

$$t_i = P(unit\ i\ used|sampled) = \frac{exp(\boldsymbol{x}_i b)P_u}{P_a + (1\text{-}P_a)exp(\boldsymbol{x}_i b)P_u}$$

$$= \frac{exp(\boldsymbol{x}_i b)\dfrac{P_u}{P_a}}{1 + (1\text{-}P_a)exp(\boldsymbol{x}_i b)\dfrac{P_u}{P_a}}$$

$$= \frac{exp(\ln(P_u) - \ln(P_a) + \boldsymbol{x}_i b)}{1 + (1\text{-}P_a)exp(\ln(P_u) - \ln(P_a) + \boldsymbol{x}_i b)}.$$

If P_a is near 0, $(1\text{-}P_a)$ can be dropped from the denominator to produce

$$t_i = \frac{exp(\ln(P_u) - \ln(P_a) + \boldsymbol{x}_i b)}{1 + exp(\ln(P_u) - \ln(P_a) + \boldsymbol{x}_i b)}\,,$$

which is reasonably accurate if $P_a < 10\%$. This expression is recognizable as a logistic regression equation with intercept equal to $\ln(P_u) - \ln(P_a) + b_0$. Under the assumed form for w^* and provided $\boldsymbol{x}_i b < 0$, regular logistic regression can be used to obtain maximum likelihood estimates of $[\ln(P_u) - \ln(P_a) + b_0]$ and all coefficients in b except b_0. An offset term in the logistic regression is not necessary in this case. Here, an RSF that is proportional to w^* can be obtained by substituting all estimated coefficients except b_0 into Equation (4).

DISCUSSION

If $\boldsymbol{x}_i b > \log(P_u/(1\text{-}P_u))$ in Equation (3), or $\boldsymbol{x}_i b > 0$ in Equation (4), $w^*(\boldsymbol{x}_i b)$ will not be a proper probability because $w^*(\boldsymbol{x}_i b) > 1$. When this occurs, some units in the population have probability of selection near or equal to 1.0, implying that the smooth functions defined in Equations (3) and (4) are poor approximations to w^* because the true w^* is not smooth. In other words, w^* in some range of the covariates is not differentiable. If this occurs, either a different form for w^* or direct maximization of the likelihood with bounds on w^* set at 0 and 1 is necessary.

Notwithstanding, $\boldsymbol{x}_i b$ will usually be less than $\log(P_u/(1\text{-}P_u))$ and 0 in Equations (3) and (4) respectively because the proportion of used units in the population will generally be small. If the proportion of used units is small, there are many more unused units in the population than used, and the true intercept parameter b_0 will be a large negative number that dominates the sum $\boldsymbol{x}_i b$. In this case, all that is required to meet the criterion on $\boldsymbol{x}_i b$ is to sample enough available units to assure accurate estimation of the true b_0. An accurate estimate of b_0 will be obtained if the ratio of the number of available units to the number of used units in the combined sample is approximately the same as the ratio of the number of unused units to the number of used units in the population. For example, if a population contains 100,000 unused units and 1000 used units, and data collection produces a sample of 500 used units, the goal should be to sample at least 500(100,000/1000) = 50,000 available units.

Unfortunately, the number of used and unused units in a population is nearly always unknown and it is almost always impossible to assess whether or not enough available units have been sampled to assure accurate estimation of b_0. Furthermore, unless P_u and P_a are known, it is not currently possible to rigorously check whether $\boldsymbol{x}_i b < \log(P_u/(1-P_u))$ or $\boldsymbol{x}_i b < 0$. Fortunately, most studies are short in duration and the proportion of units used in the population during the study period will obviously be very close to 0. In practice then, it will be sufficient to sample several orders of magnitude more available units than used units. A safe rule of thumb here is that the total number of used units should be no more than 1% of the total number of units in the population, but the restrictions $\boldsymbol{x}_i b < \log(P_u/(1-P_u))$ and $\boldsymbol{x}_i b < 0$ will be satisfied in some situations where the proportion of used units is > 1%. Research into rigorous assessment of $\boldsymbol{x}_i b < \log(P_u/(1-P_u))$ and $\boldsymbol{x}_i b < 0$ is ongoing.

When overlap occurs and the overlap units (those included in both the sample of used and sample of available units) are included in the used, but not the available sample, three estimation options are available. First, if both sampling probabilities (P_a and P_u) are known, a generalized (binomial) linear model with the link function defined in Equation (2) can be estimated. This analysis is the most flexible of the three and provides the most information because the generalized linear model allows w^* to take on any functional form and estimates of the absolute probabilities of selection are obtained. Furthermore, no constraints are placed on $\boldsymbol{x}_i b$ (only that $0 < w^*(\boldsymbol{x}_i b) < 1$). Second, if P_a, but not P_u, is known, $w^*(\boldsymbol{x}_i b)$ can be defined as in Equation (3) and logistic regression with an offset term can be used to estimate the RSF. This options should be used if P_a is substantial (> ~ 10%), but the true proportion of used units in the population should be small so that the condition $\boldsymbol{x}_i b < \log(P_u/(1-P_u))$ is satisfied. Third, if both P_a and P_u are unknown but P_a is obviously near 0, $w^*(\boldsymbol{x}_i b)$ can be defined as in Equation (4) and logistic regression can be used to estimate the RSF. While this option will be the one most often used in practice because P_a and P_u are rarely known, it should only be used if P_a is obviously small (< ~ 10%) and only when the true proportion of used units in the population is small (so that $\boldsymbol{x}_i b < 0$).

Manly et al. (2002, section 5.4) indicate that if the sample of available units was taken without replacement before the sample of used units, logistic regression can be used to estimate a resource selection function of the form defined in Equation (4), i.e.,

$$w^*(\boldsymbol{x}_i b) = exp(\boldsymbol{x}_i b) .$$

A careful inspection of the Manly et al. (2002) methods reveal that they also require the number of used units in the population to be small, and $\boldsymbol{x}_i b < 0$, but they do not require P_a to be small. When these three conditions are satisfied, the Manly et al. method and the third analysis option of this paper (Equation (4)) are for all practical purposes identical because the probability of obtaining units in the overlap is very small. If P_a is not small and overlap units are included in the sample of used units, P_a must be known or estimated in order to use logistic regression as described under analysis option two (Equation (3)). When P_a is large, the estimated RSF will differ from that obtained by the Manly et al. method because overlap units are included different samples and the resulting likelihood is different; however, the proportion of overlap observed in the samples has to be substantial for coefficients estimated by the two methods to be substantially different. In any case, the RSF values estimated by the two methods will be different due to the different assumed form for w^* when P_a is large.

LITERATURE CITED

MANLY, B. F. J., L. L. MCDONALD, D. L. THOMAS, T. L. MCDONALD, AND W. P. ERICKSON. 2002. Resource selection by animals: statistical design and analysis for field studies, 2nd edition. Kluwer Academic Publishers, Boston.

MCCULLAGH, P., AND J. A. NELDER. 1989. Generalize linear models, 2nd edition. Chapman and Hall, London.

MCCRACKEN, M. L., B. F. J. MANLY, AND M. V. HEYDEN. 1998. The use of discrete-choice models for evaluation resource selection. Journal of Agricultural, Biological, and Environmental Statistics 3: 268–279.