Ideas in Ecology and Evolution 3: 26-27, 2010
doi:10.4033/iee.2010.3.6.c

Received 17 August 2010; Accepted 28 September 2010

Commentary

Presence-only data, pseudo-absences, and other lies about habitat selection

Mark S. Boyce

M. S. Boyce (boyce@ualberta.ca), Department of Biological Sciences, University of Alberta, Edmonton T6G 2E9 Canada

Presence-only data, strictly speaking, refers to observations of the locations where animals or plants are found. By recording selected covariates to characterize those locations we can identify the domain of habitat for a species. Indeed, the distribution of covariates associated with these location data can be used to define an "envelope" of sites where an organism has been located (Pearce and Boyce 2006). Such an envelope might be mapped to show where an organism could occur presuming that we have recorded an ecologically relevant assemblage of covariates that describe the realized niche.

In practice, researchers attempt to contrast the distributions of covariates associated with a sample of an organism's recorded locations with a sample of random landscape locations. Doing so allows estimation of a Resource Selection Function (RSF) which can be used to predict which resource units (e.g., pixels) are likely to be selected by an animal. To force the data to fit the framework of logistic regression, a popular statistical method, the random landscape locations are presumed to be absences. These are sometimes called pseudo absences. In such a scheme, resource units with recorded locations will be assigned 1's and those drawn as random landscape locations will be assigned 0's. If these were true used and unused (= presence and absence) data, logistic regression could be used to estimate the probability of occupancy (or use) based on measurements of the set of covariates (MacKenzie et al 2006).

In reality, however, the sample of used locations always will be a subset of the random landscape locations. In other words, the data reflect a sample of used resource units drawn from a larger pool of resource units that were available (use/available design) and they are being forced into a logistic regression framework for statistical convenience. Unfortunately the underlying premise of two exclusive categories in a logistic regression classification is not correct. Instead there is a problem with the 0's. If the organism is located, say within a pixel, we can be sure that that location is correctly classified as a used location. But the random landscape locations where the organism was not observed might be unused because of detection bias (MacKenzie et al 2006) or because sampling intensity was insufficient (Boyce et al 2002). Therefore we have an asymmetry of errors with the 0's harbouring much greater uncertainty than the 1's.

Instead of trying to force the data into a statistical framework that is inappropriate for the data, it makes much more sense to recognize that the distribution of used locations comes directly from the distribution of random landscape locations. Indeed, this is precisely the statistical framework for applications of quantitative genetics to model natural selection (Manly 1985) that ultimately led to the development of RSFs by Lyman McDonald and Bryan Manly (Manly et al 1993). Seber (1984) developed the relevant statistical framework for this use/available design by identifying a logistic discriminant function that can be used to contrast a distribution of used resource units with those available. The RSF is the model that can be used to identify used resource units given a distribution of available resource units. If the two distributions are normally distributed, this selection function is an exponential model (Seber 1984). We can estimate coefficients for this RSF using software for logistic regression, essentially cheating the logistic regression MLE algorithm into estimating the logistic discriminant function (Johnson et al 2006). The predictive capability of this RSF can be evaluated using *k*-fold cross validation (Boyce et al 2002), thereby overcoming the inappropriate application of the

confusion matrix and related statistical procedures such as the receiver operating characteristic (ROC) and the area under the curve (AUC).

The theory is not quite so clear for the case-control design. In practice the case-control design usually involves matching a used resource unit with a random resource unit(s) drawn from within a buffer surrounding the used resource unit. This provides a distinct advantage because it restricts the domain of available resource units to a smaller set within a limited time frame, i.e., a set that is truly available for selection. Baascha et al (2010) concluded that such a case-control design produced models that were more predictive than alternative methods for estimating the RSF. However they came to this conclusion using ROCs that clearly are not an appropriate statistical method for evaluation of a use/available RSF (Boyce et al 2002). Another advantage is that the buffer size can be manipulated offering flexibility in the scale from which available resource units are drawn (Boyce 2006). Conceptually I do not see any reason that the use/available discriminant function could not be fitted using conditional logistic regression but the theory will need careful consideration.

Desrochers et al.'s (2010) idea that a time constraint will render the case-control design immune from statistical bias has some merit. But perhaps more important is the utility for identifying the context of the biological process of selection by animals. Sampling protocols can help to identify the appropriate space-time constraint to a domain over which habitat selection can occur. For example, if animal location data are being obtained by radio-telemetry, relocations might be taken at regular intervals, say each hour. By monitoring animal movement we can estimate the maximum distance that an animal is likely to move within an hour and thereby have a sound basis for the domain from which to sample available resource units.

Estimating models of habitat selection by animals does not involve presence-only data. To estimate selection we must know something about which habitats are being selected relative to what is available (Manly et al 1993, Johnson et al 2006), or alternatively we can estimate occupancy from observations of used and unused resource units (MacKenzie et al 2006). Appropriate statistical methods for estimating selection functions have been developed given the sampling scheme that generated the data, and it should not be necessary to force the data into a statistical framework inappropriate for the data.

References

Baascha, D.M., Tyrea, A.J., Millspaugh, J.J., Hygnstroma, S.E., and K.C. Vercauteren. 2010. An evaluation of three statistical methods used to model resource selection. Ecological Modelling 221:565–574. CrossRef

Boyce, M.S. 2006. Scale and resource selection functions. Diversity and Distributions 12:269–276. CrossRef

Boyce, M.S., Vernier, P.R., Nielsen, S.E., and F.K.A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 157:281–300. CrossRef

Desrochers, A., McIntire, E.J.B., Cumming, S.G., Nowak, J., and S. Sharma. 2010. False negatives—A false problem in studies of habitat selection? Ideas in Ecology and Evolution 3:20-25. CrossRef

Johnson, C.J., Nielsen, S.E., Merrill, E.H., McDonald, T.L., and M.S. Boyce. 2006. Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. Journal of Wildlife Management 70:347–357. CrossRef

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., and J.E. Hines. 2006. Occupancy estimation and modeling. Academic Press, New York, New York, USA.

Manly, B.F.J. 1985. The statistics of natural selection. Chapman and Hall, London.

Manly, B.F.J., McDonald, L.L., and D.L. Thomas. 1993. Resource selection by animals. Chapman and Hall, London.

Pearce, J.L., and M.S. Boyce. 2006. Modelling distribution and abundance with presence-only data. Journal of Applied Ecology 43:405–412. CrossRef

Seber, G.A.F. 1984. Multivariate observations. J. Wiley, New York. CrossRef