Letters

MARINE MAMMAL SCIENCE, 20(2):353–355 (April 2004)

Modeling Species-Habitat Relationships in the Marine Environment
A Comment on Hamazaki (2002)

There is a growing interest in the development of spatial models to predict marine animal distributions. However there are considerable theoretical and methodological hurdles to overcome. It is crucial that these issues be discussed plainly and openly so that we may learn from each other, and avoid common pitfalls. Hamazaki (2002) developed predictions of cetacean habitats in the mid-western North Atlantic using logistic regression models. While the work was impressive in scope—covering 13 species and applying a number of different analyses—it overlooked key methodological issues, casting doubt on the conclusions.

As is common with the development of species-habitat relationships, Hamazaki used a grid to divide the ocean into cells, making these the unit of observation. A consequence of this approach is that the resulting sample (*i.e.*, collection of grid cells) becomes inflated in size, and highly autocorrelated. However these effects were ignored, and results were presented as "highly significant" (table 2) based on a Chi-square test, in violation of the necessary independence assumption. Equally egregious was the development of logistic models with no justification for the selection of predictor variables. Understanding the relationships in one's data is accepted as fundamental to the development of any ecological model (*e.g.*, Hilborn and Mengel 1997). Hamazaki's selection of predictor variables appears to be based solely on an automated, stepwise selection approach. This is of concern because, in addition to the lack of independence among sample units, the stepwise approach is further compromised by an artificially large sample size. Although choosing predictors based on data availability is often unavoidable, their inclusion, and the selection of quadratic and interaction terms still requires justification. Selected predictors should subsequently be tested for colinearity, and the interactions between them explored in detail to ensure that all hypothesized statistical relationships are on a reasonable ecological footing (*e.g.*, Gregr and Trites 2001, Maury *et al.* 2001). The potential effects of multicolinearity can be significant (Zar 1996), and should not be ignored.

I am also concerned about the confidence expressed in the results, which was based on "effective classification rates" (which I assume were derived from classification tables). While this approach is not unreasonable, these tests are sensitive to the threshold values selected. Some discussion of the threshold values and their effects on the results would, therefore, have been valuable, particularly since this approach was subsequently used to evaluate what was termed a "sensitivity analysis" across spatial scales. In this analysis, Hamazaki generated a series of regressions using increasing grid cell size (*i.e.*, increasing spatial scale), and evaluated their performance as predictive tools using classification tables. However Hamazaki's conclusion that "96-km squares and possibly larger ... are sufficient for prediction of oceanwide cetacean habitat" is on a poor foundation without a discussion of the threshold values used, a rationale for their application, and the behavior of the predictor variables at different spatial scales. An exploration of how the predictor variables behaved as the spatial scale was increased—perhaps beyond the scale of autocorrelation—would have

provided more insight into what defines marine mammal habitat, and also come closer to a true sensitivity analysis.

Predictive models of species-habitat relationships in the marine environment (Moses and Flinn 1997, Gregr and Trites 2001, Guinet *et al.* 2001, Hamazaki 2002) have, to date, all been done on rasters (*i.e.*, grids). As discussed in Gregr and Trites (2001), this shifts the sampling unit from species observations to the grid cell. While this is convenient for analytical methods such as regression analysis, it also creates a data set that is zero-inflated (contains artificially high number of zero values), exhibits strong spatial autocorrelation, and is often effort biased.

One problem introduced by the inflated data set is "the problem of large sample size." As pointed out by Hays (1963): "Virtually any study can be made to show [statistically] significant results if one uses enough subjects regardless of how nonsensical the content may be." This effect is also recognized by Tabachnick and Fidell (2001) who describe the danger of having too much power, suggesting that rejection of the null hypothesis may be trivial if the sample size is large enough to reveal any difference whatsoever. The effect of large sample sizes in null hypothesis significance testing (NHST) appears well understood in the social sciences, although dealing with it is not (see Germano 1999).

The creation of an inflated sample size is not the only problem introduced by rasterizing the data. A related issue is the consequent spatial autocorrelation of the grid cells. Spatial autocorrelation exacerbates the pre-existing problem of non-independence among the predictor variables by adding additional structure to the data that cannot be addressed in a simple, logistic regression. In addition, the raster sample is unlikely to conform to any standard parametric distribution, casting doubt on the use of parametric statistics. This is problematic for both the selection of regression parameters and for subsequent significance tests.

Fortunately, the independence of predictors can be tested (*e.g.*, Menard 1995), and the extent of spatial autocorrelation can be measured using variograms (*e.g.*, Manly 2000). Eventually, we shall have to start including spatial correlation structures in our models to account for the autocorrelation. In the meantime, it is crucial that we recognize the problems we introduce during data rasterization, and moderate the confidence we have in our results accordingly.

While the fit of species-habitat models has been tested with classification tables, this method is extremely sensitive to the arbitrarily selected threshold value (the value at which the continuous probability distribution—generated by a logistic regression—is turned into presence-absence counts needed for classification analysis). Since the value fundamentally impacts the results, this test is considerably less than ideal. Improved methods for testing the fit of spatial models are urgently needed.

Ultimately, we are interested in the processes which drive the patterns that we see. However "because we are so clever at devising explanations of what we see, we may think we understand the system when we have not even observed it correctly" (Wiens 1989). We should, therefore, be cautious when interpreting the results of our modeling efforts, and be sure to consider a range of alternative interpretations. This is particularly true since the issues described herein are further compounded by the effects of scale, the discussion of which is well beyond the scope of this letter. In light of these issues, I think it is clear that our progress towards meaningful descriptions of marine mammal habitat will be enhanced if we share and discuss what we don't know, in addition to what we have learned.

Literature Cited

Germano, J. D. 1999. Ecology, statistics, and the art of misdiagnosis: The need for a paradigm shift. Environmental Review 7:167–190.

Gregr, E. J., and A. W. Trites. 2001. Predictions of critical habitat for five whale species in the waters of coastal British Columbia. Canadian Journal of Fisheries and Aquatic Science 58:1265–1285.

GUINET, C., L DUBROCA, M. A. LEA, S. GOLDSWORTHY, Y. CHEREL, G. DUHAMEL, F. BONADONNA AND J. DONNAY. 2001. Spatial distribution of foraging in female Antarctic fur seals (*Arctocephalus gazella*) in relation to oceanographic variables: A scale-dependent approach using geographic information systems. Marine Ecology Progress Series 219:251–264.

HAMAZAKI, T. 2002. Spatiotemporal prediction models of cetacean habitats in the mid-western North Atlantic Ocean (from Cape Hatteras, North Carolina, U.S.A. to Nova Scotia, Canada). Marine Mammal Science 18:920–939.

HAYS, W. L. 1963. Statistics. Holt, Rinehart & Winston, New York, NY.

HILBORN R., AND M. MENGEL. 1997. The ecological detective—confronting models with data. Princeton University Press, Princeton, NJ.

MANLY, B. F. J. 2000. Statistics for environmental science and management. Chapman & Hall/CRC, New York, NY.

MAURY, O., D. GASCUEL, F. MARSAC, A. FONTENEAU AND A. DE ROSA. 2001. Hierarchical interpretation of nonlinear relationships linking yellowfin tuna distribution to the environment in the Altantic Ocean. Canadian Journal of Fisheries and Aquatic Science 58:458–469.

MENARD, S. W. 1995. Applied logistic regression analysis. Sage Publications Inc., Thousand Oaks, CA.

MOSES, E., AND J. T. FINN. 1997. Using Geographic information systems to predict North Atlantic right whale (*Eubalaena glacialis*) habitat. Journal of Northwest Atlantic Fisheries Science 22:37–46.

TABACHNICK, B. G., AND L. S. FIDELL. 2001. Using multivariate statistics. Fourth Edition. Allyn and Bacon, Toronto, ON.

WIENS, J. A. 1989. Spatial scaling in ecology. Functional Ecology 3:385–397.

ZAR, J. H. 1996. Biostatistical analysis. Prentice Hall, Englewood Cliffs, NJ.

EDWARD J. GREGR, Marine Mammal Research Unit, Fisheries Centre, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; e-mail: gregr@zoology.ubc.ca.
Received 16 June 2003. Accepted 21 September 2003.