MARINE MAMMAL SCIENCE, 20(2):356–358 (April 2004)

Modeling Species-Habitat Relationship in the Marine Environment A Response to Gregr (2004)

I thank Gregr (2004) for providing valuable comments and criticisms of my article (Hamazaki 2002). This provides me an opportunity to clarify some of the methodologies I missed. I agree with Gregr that theoretical and methodological issues regarding a prediction model should be discussed plainly and openly so that we may learn from each other, and avoid common pitfalls. I also agree with all of his methodological concerns; however, I think Gregr missed the most important issues about building a prediction model: (1) the objective of building a prediction model, and (2) understanding the data source and sampling design.

A prediction model is built to estimate/predict distribution of marine mammals from a set of environmental variables. The model is based upon statistical regression (*e.g.*, multiple regression, logistic regression, general additive models) between the distribution of marine mammals and predictor/environmental variables (*e.g.*, SST). Additionally, regression models can be used to (1) estimate/compare the effects of predictor variables on distribution of marine mammals, and (2) to test hypotheses, which would enhance understandings of ecological relationships and processes that influence distribution/habitats of marine mammals.

However, using regression models for these objectives requires different sets of requirements that are not necessarily required for predictions: (1) every sampling location must have an equal chance of being surveyed (*i.e.*, random sampling), and (2) [though this is not a statistical requirement] ecological/biological relationships between marine mammal distribution and predictor variables should be theorized beforehand (Williams 1997, Guthery *et al.* 2001). Biased sampling would lead to biased estimation of the effects and significance of predictors. Statistical null hypothesis testing is an empirical testing of supporting or not supporting a particular theory, and thus null hypothesis testing without a solid theory is meaningless (Cherry 1998, Johnson 1999). Significant statistical relationship does not prove a causal relationship. It is most important that research data sampling should be designed to answer specific research objectives/questions (Cherry 1998, Johnson 1999).

Most prediction modeling studies utilize data that are not specifically designed for these objectives (*e.g.*, Gregr and Trites 2001, Hamazaki 2002). This is dangerous data dredging (Anderson *et al.* 2001, Johnson *et al.* 2001). It is meaningless and dangerous to conduct statistical analyses and interpret the results if the data are not designed to answer the research questions. This would also mean that available data determine research questions/objectives.

In my study I used sighting survey data that were collected by a systematic sampling method (*i.e.*, non-random line transect survey) to estimate abundance of species in areas *they are known to frequent* (*i.e.*, violation of requirement (1) above). [Very few sighting surveys are conducted in areas *where the species sought are known to be rare or absent.*] Predictor variables were selected for convenience of availability (*i.e.*, violation of requirement (2) above). [A prediction model is practically useless if the predictor variable data are not easily obtainable (Hamazaki 2002)]. The sighting survey also does not provide certainty whether the species are observed at their preferred environmental conditions (Hamazaki 2002). I also have not found a theory describing functional relationship between the distribution of marine mammals and the predictor variables. Thus, it is obvious that the data are not designed to investigate ecological relationships or processes of distribution/habitats of marine mammals. Even, the data are severely limited for prediction purposes (Hamazaki 2002).

In the light of the inappropriateness of using the data to investigate ecological relationships and hypothesis testing, most of Gregr's concerns related to these objectives are trivial. Inflated sample size, autocorrelation, and multicollinearity among predictor variables would lead to incorrect significance tests of the model and its parameters and inefficient model predictions (*e.g.*, incorrect confidence/prediction interval); however, those conditions would not severely bias parameter estimates and point estimates (Neter *et al.* 1990). In the model, I included the quadratic form because it is the simplest functional form simulating the Allee effects, and interactions among predictor variables because the strength of relationship between one predictor variable and the distribution of cetaceans (*i.e.*, parameter value of the predictor variable) often depends upon the state of another predictor variable. Because my objective was building a simple regression model with a set of variables that best explain the distribution of cetaceans, the use of an automated stepwise selection method, and no examination of the behavior of predictor variables in various sampling scales, are not of significant concern. I also employed the principle of Occam's razor for selection of predictor variables after the stepwise selection indicated potential predictor variables: (1) simpler predictor variables, and (2) no interaction terms without main effects. [I omitted this in the text.]

Acknowledging limitations of the models: (1) statistical confidence of the model (*i.e.*, interval estimate) and significance of predictor variables should not be trusted, and (2) some of the statistical relationships derived from the regression could be spurious, the prediction model still serves its purpose of indicating areas where the species are *more likely* to be present/sighted (*i.e.*, habitats). However, the model should not be used to investigate any ecological relationship, nor to test hypotheses. [This is the reason I intentionally did not discuss significance and meaning of predictor variables included in the model, nor precision and significance of the model, but discussed, rather, constraints originating from the data source (Hamazaki 2002).]

As Gregr argues, effectiveness/predictability of the model depends on threshold values to determine presence or absence of species. For this, I used the sighting probability, an expected sighting probability under the surveyed area (Hamazaki 2002: table 2, 3). Although this is subjective, I think this better-than-expected threshold is a reasonable threshold. [Again, I omitted this.] Most of the misclassifications are false positives (*i.e.*, predicted to be present but not sighted), which is expected because sighting surveys are conducted in areas *cetaceans are known to frequent.* Increasing the threshold (*i.e.*, fewer false positives) would increase the correct classification rate. However, because of the very low sighting probability the null model (*i.e.*, predicting no cetaceans in the survey area) would yield the higher correct classification rate (Hamazaki 2002). For choice of an appropriate scale, I used the correct classification rate as an indicator. Lack of significant change of the rate across the scales (Hamazaki 2002) indicates that reduction of sample data (*i.e.*, large scale) would not significantly alter the efficiency (prediction capability) of the model (*i.e.*, addition of data would not yield greater information). I prefer using fewer data if more data do not significantly improve results. The 96-km square is also about one day's worth of survey area, so that this can be a reasonable survey unit. While I agree with Gregr that more precise and objective criteria may be needed to evaluate fit of the prediction model, I think the objectives of the model essentially decide precision and scale (Hamazaki 2002).

I agree with Gregr about theoretical and methodological concerns in building spatial models for distribution of marine mammals, and all of Gregr's suggestions are certainly valuable. However, I am more concerned about applications of modeling techniques to data for whatever purposes, without considering whether the data are appropriate for these purposes "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). Building a spatial/temporal prediction model and investigating ecological processes determining distribution of species are different objectives and thus need different survey designs and analyses. Designing a sampling scheme that matches research objectives is the first step toward understanding of marine mammals and meaningful descriptions of their habitats.

Literature Cited

Anderson, D. R., K. P. Burnham, W. R. Gould and S. Cherry. 2001. Concerns about finding effects that are spurious. Wildlife Society Bulletin 29:311–316.

CHERRY, S. 1998. Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin 26:947–953.

GUTHERY, F. S., J. L. LUSK AND M. J. PETERSON. 2001. The fall of the null hypothesis: Liabilities and opportunities. Journal of Wildlife Management 65:379–384.

GREGR, E. J. 2004. Modeling species-habitat relationships in the marine environment: A comment on Hamazaki (2002). Marine Mammal Science 20:353–355.

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 Fishery and Aquatic Science 58:1265–1285.

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.

JOHNSON, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

JOHNSON, D. H., T. L. SHAFFER AND W. E. NEWTON. 2001. Statistics for wildlifers: how much and what kind? Wildlife Society Bulletin 29:1055–1060.

NETER, J., W. WASSERMAN AND M. H. KUNTER. 1990. Applied linear statistical models. Third edition. R. D. Irwin, Inc., Homewood, IL.

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

WILLIAMS, B. K. 1997. Logic and science in wildlife biology. Journal of Wildlife Management 61:1007–1015.

TOSHIHIDE HAMAZAKI, Alaska Department of Fish and Game, Division of Commercial Fisheries, 333 Raspberry Road, Anchorage, Alaska 99518-1599, U.S.A.; e-mail: toshihide_hamazaki@fishgame.state.ak.us. Received 27 August 2003. Accepted 21 September 2003.