USING GIS TO PREDICT MALLARD NEST STRUCTURE OCCUPANCY

John R. Fieberg, Michael C. Zicus, and Dan Hertel[1]

[1]U.S. Fish and Wildlife Service, Habitat and Populations Evaluation Team, 21932 State Highway 210, Fergus Falls, MN 56537

SUMMARY OF FINDINGS

We used the relationships described in a study of mallard nest structures to build a Geographic Information System (GIS) based model that would predict the probability of structure use by mallards. We assessed the model performance using data from a long-term study and used the assessment to illustrate a useful approach to predictive model building and validation. The model employed an existing GIS developed to aid in waterfowl management in western Minnesota. We used 3 predictors: 1) nest structure type, 2) 4 measures of the size of open water area containing the structure, and 3) a measure that described the mean aggregate visual obstruction of all residual cover during the early part of the nesting season (15 March – 20 April) in a buffer with a 1.6 km radius around each structure. We built the predictive model using the approach outlined by Harrell (2001), which is an alternative to data-based model selection methods (e.g., stepwise variable selection). We used a bootstrap procedure to obtain an unbiased measure of future predictive performance of the models that we fit. Unfortunately, we failed to produce a GIS model with much predictive power. A number of reasons related to the difficulty of predicting biological outcomes determined by constantly changing features in the landscape were likely responsible. The process we employed forced us to think about the problem rather than using a data-based selection algorithm to determine the most important variables in the model.

INTRODUCTION

Knowing which type of nest structure to use and where to deploy them in a landscape should be important to waterfowl managers. Zicus et al. (2006a) studied mallard (*Anas platyrhynchos*) nest structure occupancy in an attempt to understand how landscape features affected structure use. They were interested in the effect of 5 covariates, and their final fitted model was complex, including 3 interactions and 1 main effect. More nests were initiated as the size of the open water area where structures were deployed increased. Simultaneously, cover influence interacted with period of the nesting season such that nesting probability was positively associated with cover height and density early in the season and negatively associated with cover height and density late in the season.

Nest success in structures is generally good (Eskowich et al. 1998) with early nests having higher nest success (M. Zicus, Minnesota Department of Natural Resources, unpublished data). Consequently, hen mortality associated with renesting (Sargeant et al. 1984) would be reduced for hens nesting in structures early in the year. Further, brood and duckling survival from early-hatched nests is believed to be greater than that of later-hatched nests (e.g., Rotella and Ratti 1992, Dzus and Clark 1998, Krapu et al. 2000). These understandings led Zicus et al. (2006a) to recommend that nest structures be deployed in larger wetlands where early-season residual cover in the surrounding uplands was most abundant within 1 km of the structure. Geographic Information System (GIS) models might provide powerful tools to help waterfowl managers decide where nest structure should be placed in complex landscapes.

OBJECTIVES

- Build a GIS-based model that wildlife managers can use to help determine best placement of mallard nest structures,

- assess the model performance using data from a long-term study, and

- as a secondary objective, illustrate a useful approach to predictive model building and validation.

METHODS

We used the relationships described in a study of mallard nest structures (Zicus et al. 2006a) to build a GIS-based model that would predict the probability of structure use by mallards. The response that we were interested in modeling was the mean number of mallard ducklings (DUCKS) produced in each structure included in a long-term study of mallard nest structures (Zicus et al. 2006b). We used 3 predictors: 1) nest structure type (TYPE), 2) 4 measures of the size of open water area containing the structure (NWI, GAP, FSA03, FSA97), and 3) a measure that described the mean aggregate visual obstruction (MVOM) of all residual cover during the early part of the nesting season in a buffer with a 1.6 km radius around each structure.

Data used to build the model

We began with a GIS developed to aid in waterfowl management in western Minnesota (D. Hertel, unpublished data). Classified Landsat Thematic Mapper data from 2000 and 2001 was used to estimate the area of each habitat class within buffers (1.6 km radius) around each nest structure.

The following variables were included in the model:

DUCKS. – We determined the mean number of ducklings from 110 nest structures across the entire nesting season from 1996 – 2003 (M. Zicus, unpublished data).

TYPE. – We considered 2 types of cylindrical nest structures, those having either a single or a double cylinder (Zicus et al. 2006a).

Open water area measures. – Different measures of the size of the open water area containing the structure were determined to compare model performance with different data sources. These measures were from: 1) open water polygons in National Wetland Inventory data (i.e., NWI; D. Hertel, unpublished data), 2) areas classified as open water in MN-GAP land cover data (i.e., GAP; Minnesota Department of Natural Resources 2004, U. S. Geological Survey 1989), 3) open water areas digitized from 2003 Farm Services Agency (FSA) aerial photography (i.e., FSA03; M. Zicus, unpublished data), and 4) open water areas digitized 1997 FSA aerial photography (i.e., FSA97; Zicus et al 2006a). The distribution of the NWI water data was highly skewed. As a result, we expected a few data points with extreme values (e.g., >100 ha) to have substantial influence on the model fit. Therefore, we also consider log(NWI + 0.1) which had a more bell-shaped distribution. Both NWI and GAP data are readily available for large areas of western Minnesota, whereas FSA97 and FSA03 data were included here to determine the potential gain in predictive power that might be obtained if efforts were made to obtain more up-to-date measures of open water.

MVOM. – We created a variable for the mean aggregate visual obstruction measurement (MVOM) for 15 March – 20 April for each buffer around each structure (D. Hertel, unpublished data). First, each 28 m x 28 m GIS cell within a particular habitat class in the buffer was assigned a habitat-specific VOM (Table 1). Next, a weighted VOM was calculated for each cell in a particular habitat class by multiplying the area of that habitat class in the buffer by the habitat-specific VOM. A mean aggregate visual obstruction measurement (MVOM) was then

calculated for all cells in the buffer by summing the weighted VOMs across all habitat classes in the buffer and dividing by the total area of the buffer.

Modeling

We built predictive models using the approach outlined by Harrell (2001). We first determined a reasonable degree of model complexity using guidelines based on our sample size. This approach can be summarized as "determine the number of degrees of freedom (df) that can be spent, and then spend them without any further model simplification." Harrell suggested a minimum of 10 – 20 observations per parameter considered, including those that account for potential non-linear effects. Burnham and Anderson (1998) suggested a similar liberal rule of 10 observations per predictor. Consequently, we believed 5-10 parameters to be a maximum for the 110 structures that we observed.

We used Spearman's ρ^2 (i.e., between response and predictors) to help determine how to apportion the df among the available predictors (e.g., to account for potential non-linearities) (Harrell 2001). Spearman's ρ^2 is a generalization of the rank correlation between two variables that can account for nonmonotonic relationships (e.g., using quadratic ranks) (Harrell 2001:127). We included all variables for which we examined ρ^2 in the model (i.e., ρ^2 was used only to determine the degree of non-linearity in the model). These steps defined an *a priori* full model from which we made our inferences; thereby avoiding problems associated with model selection algorithms (e.g., over fit models that predict new data poorly and biased p-values and confidence intervals arising from models selected using data-based selection procedures).

We used a bootstrap procedure to obtain an unbiased measure of future predictive performance of the models that we fit (Harrell 2001). We fit the model to 1,000 bootstrapped data sets, and the fitted parameters were used to calculate predicted values for all observations in the original

dataset (as well as the bootstrap data set). We then calculated two R^2 values for each bootstrap replication: 1) using the original data and predicted values from the bootstrap model fit and 2) using the bootstrap data and the predicted values from the bootstrap model fit. The difference between these two values is an estimate of "optimism" (i.e., resulting from fitting and "testing" the model on the same dataset). A final adjusted R^2 value was then determined by subtracting the mean "optimism" from the R^2 obtained from the original fit of the model to the full dataset. Bootstrap calculations were carried out using functions in the Design library of the R computing package (Harrell 2001, R Core Development Team 2005). We also calculated the usual adjusted R^2.

RESULTS

Model complexity

Values of Spearman's ρ^2 indicated that both TYPE and MVOMs had less potential for explaining variation in DUCKS than open water area (Figure 1). Consequently, we assumed the MVOM effect was linear (i.e., a single df was used to model the relationship between MVOMs and DUCKS). The relatively greater values of Spearman's ρ^2 for open water area and previous work (Zicus et al. 2006a) suggested that more dfs should be spent to model the effect of open water area. Values of Spearman's ρ^2 were considerably higher for the digitized water measures (FSA03 and FSA97) than either NWI or GAP measures of open water.

Two models were fit using digitized water data (FSA03 and FSA97):

DUCKS = TYPE + MVOM + water (using a linear spline with 2 df), and (1)

DUCKS = TYPE + MVOM + water (using a restricted cubic spline with 2 df). (2)

Model (1) used a single knot (i.e., the location where the slope was assumed to change), while model (2) used 3 knots (2 of these were located at the boundary of the data; the fit of a restricted cubic spline is constrained to be linear outside the range of the boundary knots). The medians of non-zero observations (3.66 and 3.14 for FSA03 and the FSA97 data, respectively) were chosen as the knot location for the linear spline. Knots for the cubic spline used the 10^{th}, 50^{th}, and 90^{th} percentiles of the data.

The GAP data only had 6 observations that were >0 and were not considered further. Given the low values of Spearman's ρ^2 for the NWI water data, we considered a model that assumed the effect of open water area was linear. In addition, we examined a model with a 2 dfs restricted cubic spline with knot locations again determined using the 10^{th}, 50^{th}, and 90^{th} percentiles of the data.

Estimates of predictive power

Models that used FSA03 and FSA97 water data performed considerably better than models using the NWI or GAP water data (Table 2). However, none of the models performed particularly well. The model using the FSA97 data had an R^2 of 0.14, suggesting that the open water area measured in Zicus et al. (2006a) along with structure type and MVOM values explained 14% of the variation in mean duckling production per structure. However, bootstrap validation suggested this model would perform considerably worse when applied to new data (i.e., it would explain only 6% of the variation). By comparison, R^2 measures for models using the NWI data were all less than 5% and their adjusted measures were negative, suggesting that the grand mean might predict new data better than the fitted model.

TYPE and MVOM values had p-values considerably >0.05 in all of the models, suggesting that they were not associated DUCKS (see also exploratory plots with smoothing lines; Figure 2). These results suggest that the MVOM values are not likely to be useful for predicting the mean duckling production (across all periods and years) in nesting structures and that the available measures of open water area (NWI and GAP) are of questionable value for modeling duckling production.

DISCUSSION

Models having strong predictive ability are often difficult to construct (Steyerberg et al. 2001, Ambler et al. 2002, Steyerberg et al. 2003). There are a number of reasons why our efforts may have failed to produce a GIS model with much predictive power. First, mean visual obstruction measurements (MVOM) within 1 km of each structure may not accurately reflect the importance of surrounding cover. In particular, the height and density of cover in individual buffers having the same land use could actually differ markedly. Second, while Zicus et al. (2006a) recommended making structure placement decisions using early spring landscape conditions (as described by aggregate MVOMs in the buffer), their recommendations were intended to encourage production of young early in the season and not necessarily the maximum production of young across the entire nesting season. Zicus et al. (2006a) found that occupancy rates increased with VOM measurements early in the nesting season and decreased with VOMs later in the nesting season. Given the time-varying effect of VOM on occupancy rates, it was not surprising to discover that MVOM was unrelated to season-long duckling production. Lastly, although cover and water body size both vary temporally, we were forced to use measurements of these variables from a single year. The relationship between these habitat measurements and the average productivity of structures (across the 8 years of the study) may be much weaker than the relationship between habitat covariates and productivity in any given year.

The question as to how much predictive power a model would need to have in order to be useful is difficult to answer. Regardless, the models using either NWI or GAP measures of open water had essentially no predictive power, and a better measure of open water would be needed to produce a model with even low predictive ability. FSA97 open water values produced the model with the most predictive ability, but even this was low, perhaps because water conditions had changed significantly between 1997 and 2003. Identifying specific locations for management actions such as nest structures will be difficult when the desired biological outcomes are determined by features in the landscape that are constantly changing. A sensible strategy for structure placement and management would be to place structures in larger wetlands (>10 acres) where early-season residual cover in the surrounding uplands is most abundant (Zicus 2006a; Minnesota Department of Natural Resources. 2006. Using cylindrical nest structures to increase mallard nest success. Unpublished pamphlet.). This should reduce the number of structures that never get used as 19 of 20 structures that were not used during the 8-year study were deployed in open water areas <2 acres in size (M. Zicus, unpublished data). In addition, we recommend that managers continue to collect data on structure use as well as habitat measurements surrounding the structure (e.g., cover types, wetland size) so that we might refine our models in the future.

Despite the poor predictability of the models considered, we believe the general modeling approach is a useful alternative to data-based model selection methods (e.g., stepwise variable selection). Harrell (2001:56-57) provides 7 disadvantages of stepwise selection methods (repeated verbatim below):

1. It yields R^2 values that are biased high.

2. The ordinary *F* and χ^2 test statistics do not have the claimed distribution. Variable selection is based on methods (e.g., *F* tests for nested models) that were intended to be used to test only prespecified hypotheses.
3. The method yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow.
4. It yields P-values that are too small (i.e., there are several multiple comparison problems) and that do not have the proper meaning, and the proper correction for them is a very difficult problem.
5. It provides regression coefficients that are biased high in absolute value and need shrinkage. Even if only a single predictor were being analyzed and one only reported the regression coefficient for that predictor if its association with *Y* were "statistically significant," the estimate of the regression coefficient $\hat{\beta}$ is biased (too large in absolute value). To put this in symbols for the case where we obtain a positive association ($\hat{\beta}$ >0), E($\hat{\beta}$ | *P* < 0.05, $\hat{\beta}$ > 0) > β.
6. Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.
7. It allows us to not think about the problem.

Wildlife biologists have become familiar with problems associated with stepwise selection methods due to the popular book by Burnham and Anderson (2002) on model averaging and multi-model inference. As a result, model averaging and multi-model inference using AIC weights (Burnham and Anderson 2002) have become exceedingly prevalent in the wildlife literature. Unfortunately, few alternatives to AIC model averaging have been presented in applied ecology/wildlife journals (Guthery et al. 2005), and therefore model averaging is applied routinely without critical thinking. We would argue that approaches that utilize a full model with candidate predictors chosen based on subject matter considerations will often provide a viable alternative to model averaging/multi-model inference. The former approach offers several

advantages over the AIC-based model-averaging paradigm. For example, more time can be spent on diagnostics and model validation since a single model is considered rather than a suite of candidate models. In addition, if interest lies in estimation (rather than prediction), calculation of valid confidence intervals is straightforward (estimates of regression coefficients and σ^2 are not biased from considering multiple models or model reduction) (Harrell 2001, Ambler 2002).

The benefits of using a full model for inference are likely to be greatest when the effective sample size is >10 – 20 times the number of candidate predictors (Harrell 2001, Ambler 2002). For problems where the ratio of effective sample size to number of predictors is smaller, we recommend first trying to eliminate variables that do not have strong biological support (e.g., based on prior studies). This process is advantageous because it forces the researcher to think about the problem rather than using a data-based selection algorithm to determine the most important variables. In addition, it is generally beneficial to eliminate redundant variables, variables with lots of missing values, and variables that have very narrow distributions (Harrell 2001). If the number of remaining predictors is still >10 – 20 times the effective sample size, model averaging or other methods of shrinkage (e.g., penalized estimation or lasso) may offer improved predictions (Harrell 2001, Ambler 2002).

ACKNOWLEDGMENTS

Dave Rave collected much of the nest structure data that we used to build and validate the model. We also thank Rex Johnson for his suggestion regarding source data for the GIS model.

LITERATURE CITED

AMBLER, G., A. R. BRADY, AND P. ROYSTON. 2002. Simplifying a prognostic model: A simulation study based on clinical data. Statistics in Medicine 21:3803-3822.

BURNHAM, K. P, AND D. R ANDERSON. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

DZUS, E. H., AND R. G. CLARK. 1998. Brood survival and recruitment of mallards in relation to wetland density and hatching date. Auk 115:311-318.

ESKOWICH, K., D. MCKINNON, G. BREWSTER, AND K. BELCHER. 1998. Preference and use of nest baskets and nest tunnels by mallards in the parkland of Saskatchewan. Wildlife Society Bulletin 26:881-885.

GUTHERY, F. S., L. A. BRENNAN, M. J. PETERSON, AND J. J. LUSK. 2005. Information theory in wildlife science: Critique and viewpoint. Journal of Wildlife Management 69:457-465.

HARRELL, F. E., JR. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York, New York.

KRAPU, G. L., P. J. PIETZ, D. A. BRANDT, AND R. R. COX, JR. 2000. Factors limiting mallard brood survival in prairie pothole landscapes. Journal of Wildlife Management 64:553-561.

MINNESOTA DEPARTMENT OF NATURAL RESOURCES. 2004. The Minnesota Department of Natural resources data deli. Minnesota Department of Natural Resources, St. Paul, Minnesota, USA. http://maps.dnr.state.mn.us/deli/.

R CORE DEVELOPMENT TEAM. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org.

ROTELLA, J. J., AND J. T. RATTI. 1992. Mallard brood survival and wetland habitat conditions in southwestern Manitoba. Journal of Wildlife Management 56:499-507.

SARGEANT, A. B., S. H. ALLEN, R. T. EBERHARDT. 1984. Red fox predation on breeding ducks in midcontinent North America. Wildlife Monographs 89:1-41.

STEYERBERG, E. W., F. H. HARRELL JR., G. J. BORSBOOM, M. J. EIJKEMANS, Y. VERGOUWE, AND J. D. HABBEMA. 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology 54:774-781.

STEYERBERG, E. W., S. E. BLEEKER, H. A. MOLL, D. E. GROBBEE, AND K. G. MOONS. 2003. Interval and external validation of predictive models: A simulation study of bias and precision in small samples. Journal of Clinical Epidemiology 56:441-447.

U. S. GEOLOGICAL SURVEY. 1989. Gap Analysis Program. U. S. Geological Survey Biological Resources Division, Moscow, Idaho, USA. http://www.gap.uidaho.edu/.

ZICUS, M. C., D. P. RAVE, A. DAS, M. R. RIGGS, AND M. L. BUITENWERF. 2006a. Influence of land use on mallard nest structure occupancy. Journal of Wildlife Management 70:in press.

ZICUS, M. C., D. P. RAVE, AND J. R. FIEBERG. 2006b. Cost effectiveness of single- vs. double-cylinder over-water nest structures. Wildlife Society Bulletin 34:in press.

Table 1. Land use cover types and source of visual obstruction measurements (VOM) used to estimate mean visual obstruction measurements (MVOMs) in the GIS model.

GIS model		Source data		
Cover type	VOM (dm)[a]	Cover type	VOM (dm)	Reference
Grassland	1.16	CRP grass	1.30	Zicus[d]
		WMA grass	1.02	Zicus
		WPA grass	0.86	Zicus
		Other grass	0.86	Zicus
Cropland	0.001	Cropland[b]	0.001	Mack 1991
Hayland	0.80	Hayland	0.80	Mack 1991
Right-of-way	0.75	Gravel township road	0.71	Zicus
		Gravel county road	0.40	Zicus
		Gravel CSAH[c]	0.40	Zicus
		Paved CSAH	0.65	Zicus
		State highway	0.41	Zicus
		Railroad	1.60	Zicus
Woodland	1.70	Woodland	1.70	Mack 1991
Odd areas	1.70	Odd areas	1.70	Mack 1991
Vegetated wetlands	0.67	Seasonal	1.00	Mack 1991
		Semi-permanent	2.00	Mack 1991
		Temporary	0.50	Mack 1991
		Permanent	1.00	Mack 1991
Open water/barren	0.00	Open water/barren	0.00	Mack 1991

[a]Visual obstruction measurement corresponding to residual conditions in early spring (15 March – 20 April). Values are weighted by the area of the various source types occurring in western Minnesota.

[b]Mack (1991) presents values for many types of cropland. The value for fall-plowed cropland was used.

[c]CASH = county state aid highway.

[d]VOM is the mean value for 1997-1999 based on unpublished data collected as part of Zicus et al. (2006a).

Table 2. Measures of future predictive accuracy of GIS models predicting average duckling production from 110 nest structures in Grant County Minnesota, 1997 – 2003.

Model[a]	R^2		
	Original	Adjusted (from linear regression)	Adjusted (bootstrap)
FSA03, lsp	0.087	0.052	0.009
FSA03, rcs	0.084	0.050	0.009
FSA97, lsp	0.138	0.105	0.061
FSA97, rcs	0.134	0.102	0.056
NWI, linear	0.024	-0.013	-0.042
NWI, rcs	0.042	0.006	-0.045
Log(NWI), linear	0.027	0.000	-0.036
Log(NWI), rcs	0.053	0.017	-0.031

[a]lsp = linear spline model with 1 knot (2 dfs); rcs = restricted cubic spline model with 2 knots (3 dfs).

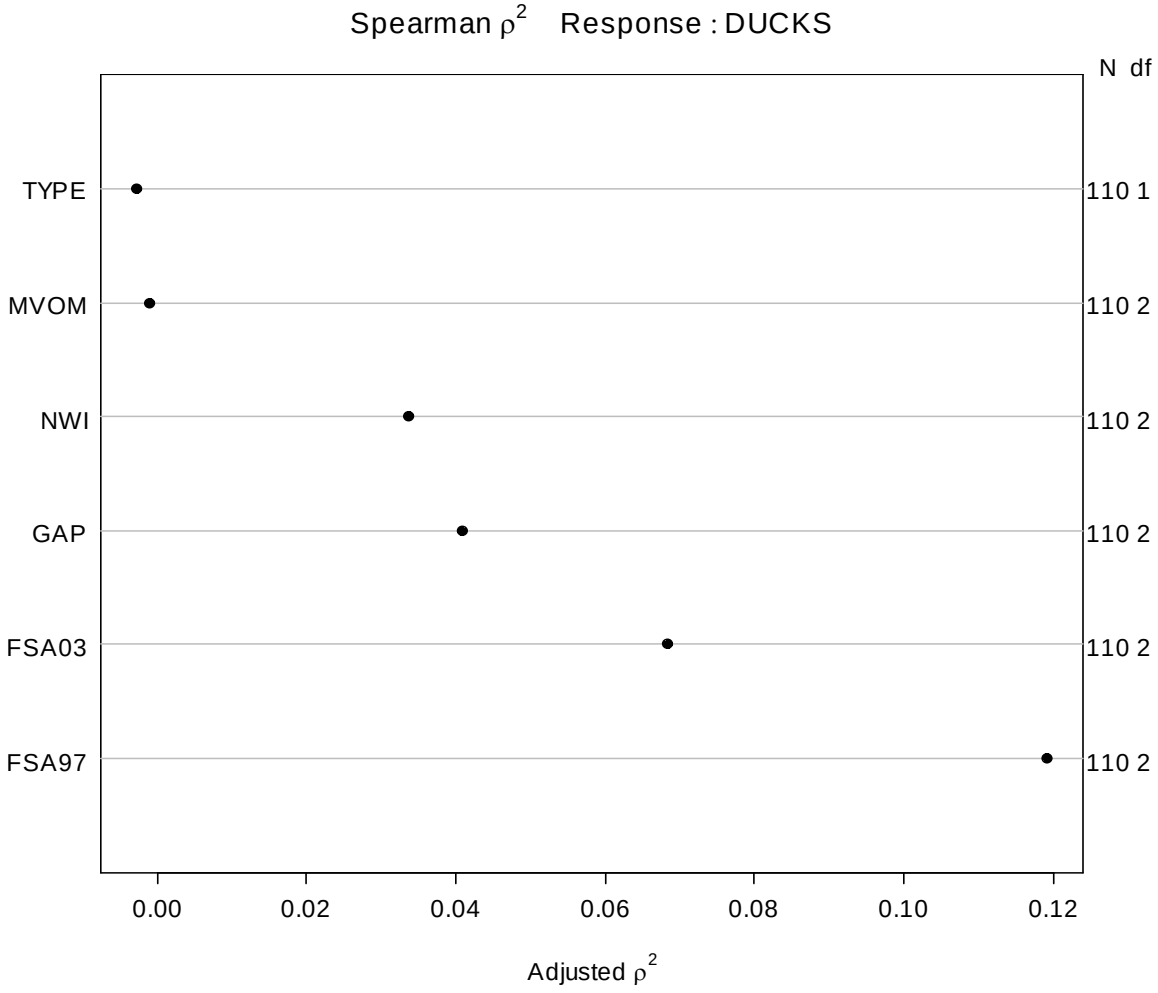


Figure 1. Spearman's ρ^2 indicating the strength of the relationship between mean ducking production (DUCKS) and each predictor variable (TYPE = indicator variable for structure type, MVOM measures, NWI open water measure, GAP open water measure, FSA03 open water measure, FSA97 open water measure).

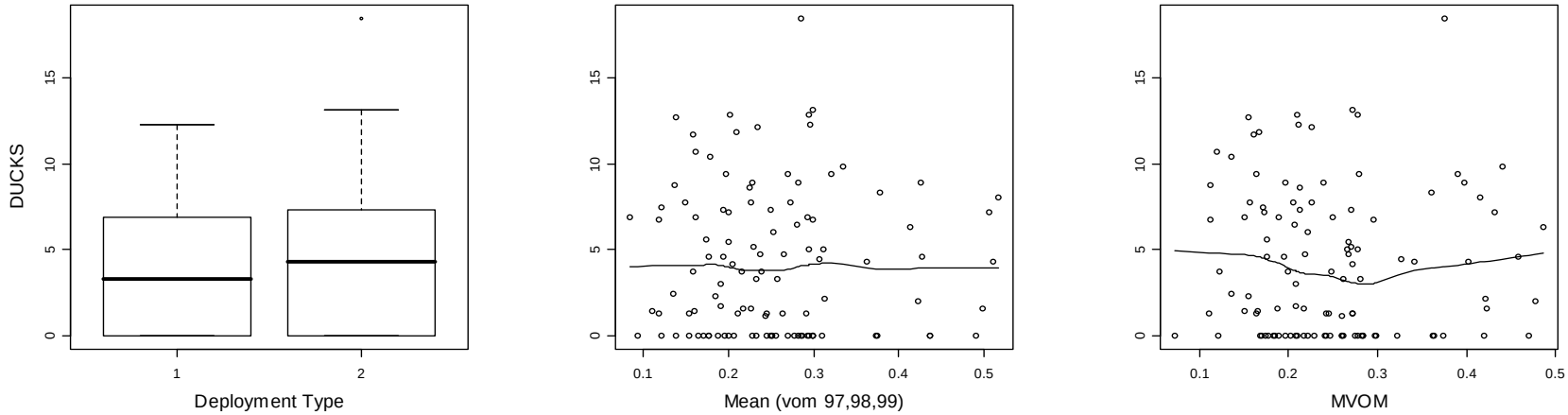
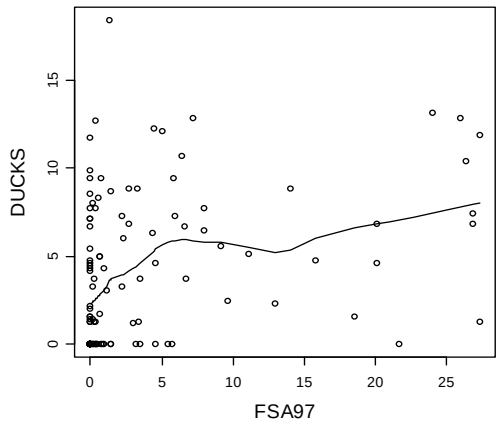
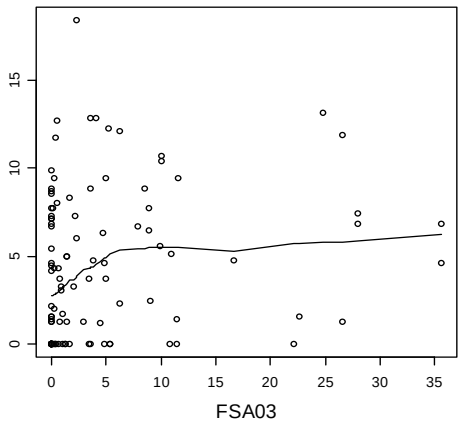
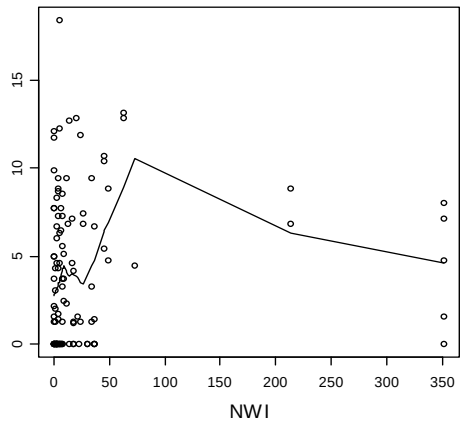
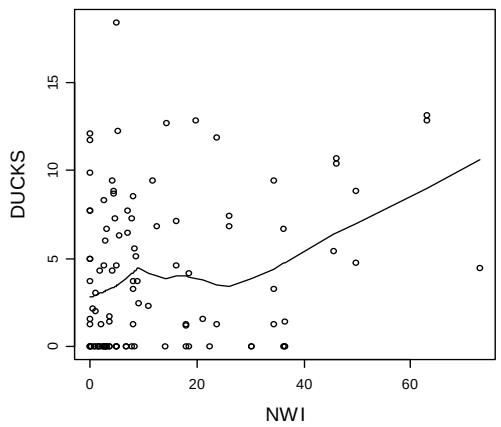


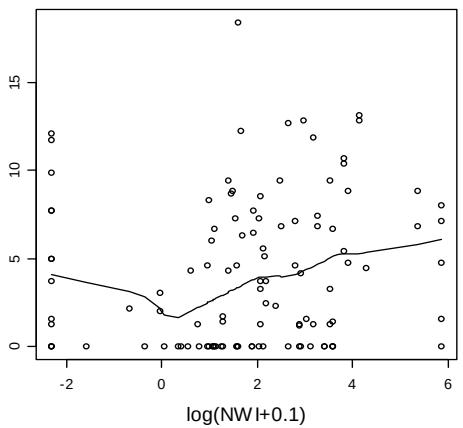
Figure 2. Exploratory plots of mean duckling production/year for each structure versus structure type, mean VOM measures across 1997-1999 (M. Zicus, unpublished data), and MVOM (D. Hertel, unpublished data). Lines represent smooth curves estimated using locally weighted regression via the lowess function in R.


DUCKS
FSA97


FSA03


NWI


DUCKS
NWI


log(NWI+0.1)

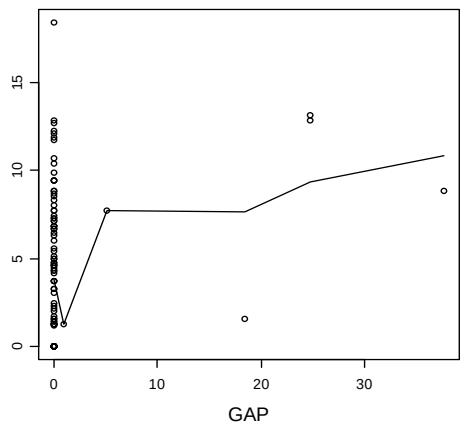

GAP

Figure 3. Exploratory plots of mean duckling production/year/structure versus FAS97 open water, FSA03 open water, NWI open water (all values), NWI open water (only values < 100), log(NWI + 0.1), and GAP open water. Lines represent smooth curves estimated using locally weighted regression via the lowess function in R.