METHODS

Statistical analysis

We analyzed the relationship of fountain darters to the potential predictor variables using multinomial logit regression, a generalized linear model (GLMs) which allow various distributions for the response and error terms in the model (Agresti, 2007). The multinomial logit regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logit regression is an extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership (Starkweather & Moske, 2011).

Based on our field experience, we categorized fountain darter density (D; number of fountain darters per m^2) into four categories (K): 1 (poor; $0 \leq D \leq 5$; 393 observations), 2 (fair; $5 < D \leq 15$; 204 obs.), 3 (good; $15 < D \leq 30$; 124 obs.), and 4 (very good; $D > 30$; 74 obs.), where mean D = 11.22 and SD (standard deviation) = 14.92. Since any of the four categories could have occurred between each observation, we assumed that fountain darter density categories did not tend to occur in any particular order and that the categories were strictly nominal. We defined the density category as the response Y_i, recorded in the location i, with $Y_i = K$. We assumed a multinomial distribution for the response Y_i with class probabilities $P(Y_i = K)$. The model has the form:

$$P(Y_i = 1) = \frac{exp(\alpha_1 + \beta_1 X_i)}{1 + \sum_{k=1}^{3}[exp(\alpha_k + \beta_k X_i)]} \quad (1)$$

$$P(Y_i = 2) = \frac{exp(\alpha_2 + \beta_2 X_i)}{1 + \sum_{k=1}^{3}[exp(\alpha_k + \beta_k X_i)]} \quad (2)$$

$$P(Y_i = 3) = \frac{exp(\alpha_3 + \beta_3 X_i)}{1 + \sum_{k=1}^{3}[exp(\alpha_k + \beta_k X_i)]} \quad (3)$$

$$P(Y_i = 4) = \frac{1}{1 + \sum_{k=1}^{3}[exp(\alpha_k + \beta_k X_i)]} \quad (4)$$

where parameter vectors α_k and β_k relate to category *k*, the vector X_i is a row of the design matrix containing independent environmental variables. Note that X_i containing only Y_i is the minimal model fitted to estimate the density probabilities. We used SAS 9.2 (SAS Institute Inc., 2008) to fit the model. We tried to fit the model to maximize the likelihood from a multinomial distribution subject to the constraints of equations 1 to 4. Having fitted the model, the density probabilities of each category in the location *i* can be calculated.

Model selection

We selected the best model by removing insignificant terms one at time and re-estimating the model (Agresti, 2007) until the Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978) could not be lowered further. We conducted model selection using SAS 9.2 (SAS Institute Inc., 2008).

Model evaluation

We evaluated the reliability and validity of our final model based on the Area Under the Curve (AUC) as fair (0.50 < AUC = 0.75), good (0.75 < AUC = 0.92), very good (0.92 < AUC = 0.97), or excellent (0.97 < AUC = 1.00) (Hosmer & Lemeshow, 2000). We computed the AUC for all six comparison pairs (e.g. Y_i = 1 vs. Y_i = 2) and averaged the result (Hand & Till, 2001). We conducted model evaluation using the pROC package (Robin *et al.*, 2011) in R 2.14.1 (R Development Core Team, 2006).

RESULTS

Results indicated that AIC reached its minimums (AIC = 1,583.038; model f; Table 1) once five variables were removed: Green_Algae, VegCover, VegHeigh, Velocity and DO. The final model included the constant and 12 variables (Table 2) with the mean AUC of 0.7972 (Table 3), which is considered good. Although some variables were not significant for certain categories ($P > 0.05$), all variables included in the final model were significant overall (Table 2). Pairwise AUC scores, which reflect the ability to discern between assignments to alternative categories, were generally higher for those categories that had some degree of separation (e.g., 1 vs 4) (Table 5). That is, as one would expect, the model had highest discerning power for more distinct categories.

Table 1 Variable selection process. Variables once removed were not returned to the model. The minimum value of AIC is in bold

Model ID	Variable removed	AIC
a	None	1608.649
b	Green_Algae	1602.651
c	VegCover	1596.703
d	VegHeight	1591.225
e	Velocity	1586.141
f	DO	**1583.038**
g	Temperature	1583.471

Table 2 Variables included in the best model according to the AIC criteria (model f)

Variable	Overall *P*-value	Category 2		Category 3		Category 4	
		Estimated coefficient	*P*-value	Estimated coefficient	*P*-value	Estimated coefficient	*P*-value
Constant	–	7.7584	0.7452	9.5793	0.0004	2.1008	<0.0001
Bryophytes	<0.0001	4.3455	0.0566	4.2244	0.0021	9.8051	0.0006
Cabomba	<0.0001	4.4947	0.0009	3.3249	<0.0001	8.6736	<0.0001
Ceratopteris	0.0329	3.4015	0.0078	1.8596	0.0181	-5.4335	0.0854
Fil_Algae	<0.0001	6.0201	0.0018	4.7125	0.0025	12.1659	0.0019
Hygrophila	<0.0001	3.5061	<0.0001	2.8677	<0.0001	6.6600	0.0007
Ludwigia	<0.0001	3.9065	<0.0001	3.8657	<0.0001	8.1887	<0.0001
Sagittaria	0.0126	2.2736	0.0054	1.2025	0.0029	5.7577	0.1088
Vallisneria	0.0005	3.0407	0.4134	1.2385	<0.0001	6.8683	<0.0001
With_Bryo	<0.0001	1.8536	0.0995	1.9350	0.0002	2.8385	<0.0001
Depth	0.0326	-0.3647	0.0002	-0.3881	<0.0001	-0.0018	<0.0001
Cond	0.0483	-0.0018	0.7300	-0.0009	0.1770	0.0002	0.0188
pH	<0.0001	-1.4116	0.3130	-1.7139	0.0143	-1.6568	0.0518

Table 3 Pairwise AUC scores for all combinations of density categories(?) and mean AUC

AUC	1 vs 2	1 vs 3	1 vs 4	2 vs 3	2 vs 4	3 vs 4	Mean
Value	0.7728	0.8596	0.937	0.636	0.805	0.7728	0.7972