Environmental Toxicology and Chemistry, Vol. 21, No. 6, pp. 1168–1175, 2002
Printed in the USA
0730-7268/02 $9.00 + .00

PREDICTING LEVELS OF STRESS FROM BIOLOGICAL ASSESSMENT DATA: EMPIRICAL MODELS FROM THE EASTERN CORN BELT PLAINS, OHIO, USA

SUSAN B. NORTON,*† SUSAN M. CORMIER,‡ MARC SMITH,§ R. CHRISTIAN JONES,|| and MARY SCHUBAUER-BERIGAN#
†U.S. Environmental Protection Agency, National Center for Environmental Assessment, Washington, DC 20460
‡U.S. Environmental Protection Agency, National Exposure Research Laboratory, Cincinnati, Ohio 45268
§Ohio Environmental Protection Agency, Ecological Assessment Section, Groveport, Ohio 43125, USA
||George Mason University, Department of Environmental Science and Policy, Fairfax, Virginia 22030, USA
#National Institute for Occupational Safety and Health, Department of Health and Human Services, Cincinnati, Ohio 45213, USA

(*Received* 31 *January* 2001; *Accepted* 25 *November* 2001)

Abstract—Interest is increasing in using biological community data to provide information on the specific types of anthropogenic influences impacting streams. We built empirical models that predict the level of six different types of stress with fish and benthic macroinvertebrate data as explanatory variables. Significant models were found for six stressor factors: stream corridor structure; siltation; total suspended solids (TSS), biochemical oxygen demand (BOD), and iron (Fe); chemical oxygen demand (COD) and BOD; zinc (Zn) and lead (Pb); and nitrate and nitrite (NO_x) and phosphorus (P). Model R^2 values were lowest for the siltation factor and highest for TSS, BOD, and Fe. Model R^2 values increased when spatial relationships were incorporated into the model. The models generally performed well when applied to a random subset of the data. Performance was more mixed when models were applied to data collected from a previous time period, perhaps because of a change in the spatial structure of these systems. These models may provide a useful indication of the levels of different stresses impacting stream reaches in the Eastern Corn Belt Plains ecoregion of Ohio, USA. More generally, the models provide additional evidence that biological communities can serve as useful indicators of the types of anthropogenic stress impacting aquatic systems.

Keywords—Multivariate regression Spatial autocorrelation Streams Macroinvertebrates Fish

INTRODUCTION

Biological assessment is becoming an increasingly popular tool in the evaluation of stream ecosystem integrity. However, little progress has been made to date in developing tools to relate assessment results to specific stressors. This paper continues the investigation of the feasibility of using fish and benthic macroinvertebrate community structure to distinguish among major types and degrees of anthropogenic stressors in the Eastern Corn Belt Plains ecoregion of Ohio, USA.

This paper builds on a previous effort that constructed a data set of spatially and temporally matched stressor and response data, reduced the stressor data to six orthogonal factors, and explored the ability of the biological community to discriminate among the different types and degrees of stress [1]. That study found that biological variables could significantly distinguish higher and lower quality sites classified on the basis of six different types of stress: quality of stream corridor structure; degree of siltation; total suspended solids (TSS), iron (Fe), and biochemical oxygen demand (BOD); chemical oxygen demand (COD) and BOD; lead (Pb) and zinc (Zn); and nitrate and nitrite (NO_x) and phosphorus (P). Functions based on biological variables could also discriminate between sites having different predominant stressors (12 of 15 pairwise combinations).

The current effort investigated the feasibility of building multiple linear regression models that predict the degree of different types of stress based on characteristics of the biological community. These predictions may provide useful information for identifying which stressors should be the subject of management action. In addition, the results can be used to evaluate whether stream communities respond in distinctive ways to different types of stress. If communities respond in consistently different ways, then distinctive models should be selected for different types of stress.

MATERIALS AND METHODS

Data set and preliminary analyses

The Eastern Corn Belt Plains ecoregion of Ohio was selected as a study location to take advantage of the large amount of biological monitoring data that was collected in a consistent manner and made available by the Ohio Environmental Protection Agency [1]. The study was confined to one ecoregion to minimize some of the natural physical, biological, and geological factors that may confound responses to stressors.

The data set contained spatially and temporally matched descriptors of fish and macroinvertebrate community structure, and variables associated with potential stressors, including instream chemistry and habitat. The complete data set encompassed the years 1988 to 1994 and included 179 sites, 42 biological variables, 18 variables associated with stressors, and the stream gradient and drainage area size associated with each sampling location. Descriptive statistics for the data set are provided in Norton et al.[1] and Norton [2].

Before analysis, the variables were transformed to near normality on the basis of visual inspection of the transformation

* To whom correspondence may be addressed (norton.susan@epa.gov).

Presented at the 20th Annual Meeting, Society of Environmental Toxicology and Chemistry, Philadelphia, Pennsylvania, November 14–18, 1999.

The views expressed in this paper are those of the author and do not necessarily reflect the views and policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

series q-q plots and results of the Shapiro–Wilkes test [3–5]. In addition, we identified variables that were significantly correlated ($p = 0.05$) with drainage area. To minimize the influence of this covariate, we fit linear and quadratic regression models to the identified variables and used the model residuals as replacement variables [1,2]. Variables that proved important in the models are shown in Appendix 1.

The 18 stressor variables and stream gradient were reduced to a set of six factors by using principal components analysis with varimax rotation in SAS® [1,4]. Stream corridor structure (factor 1) was highly correlated with ordinal scores for channel, cover, the riparian zone, and pool depth; higher scores along this factor correspond to higher quality sites (e.g., more sinuous reaches). Siltation (factor 2) was correlated with ordinal scores for riffle quality, substrate quality, and embeddedness; high scores along this factor correspond to higher quality sites (i.e., less siltation). The last four factors were highly correlated with stream chemistry variables: TSS, Fe, and BOD (factor 3), COD and BOD (factor 4); Zn and Pb (factor 5); and N and P (factor 6). High scores along the stream chemistry factors correspond to lower quality sites (i.e., higher chemical concentrations).

Model fitting

The purpose of the models was to predict the site score along each of the six factors representing different types of stress. Multiple linear regression was used as the baseline modeling approach (base models)

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i \tag{1}$$

where Y_i are the values of the stressor factor being modeled; $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters (estimated by $B_0, B_1, \ldots, B_{p-1}$); $X_{i1}, \ldots, X_{i,\, p-1}$ are the values of the explanatory (i.e., biological) variables; and ε_i are errors that are independent and normally distributed $N(0, F^2)$, estimated by the model residuals, e_i

$$i = 1, \ldots, n$$

An exploratory evaluation of the models (analysis not shown) indicated that model residuals located adjacent to one another on the same stream were significantly correlated. This result was not surprising; the issue of spatial and temporal correlation has been studied extensively in time series modeling, economics, and geography [6–9], and distance decay functions are commonly used in mechanistic models of water quality [9,10]. Spatial correlation of residuals can result in underestimates of the both variance of the error term of the regression model and the standard deviation of the estimated regression coefficients. As a result, the coefficient of multiple determination (R^2) can be overestimated, and variables can be concluded to be significant when in reality they are not [11]. For this reason, we investigated approaches to mitigate the correlation.

Three approaches for reducing the spatial correlation in model residuals were pursued. Each modeling approach was pursued for all six stressor factors and the Durbin–Watson test was used to evaluate whether the spatial correlation of the residuals was significant [6]. The first approach, the Cochrane–Orcutt procedure, uses the correlation between adjacent values to remove the influence of the spatial correlation [11]. Because the spatial correlation is removed before the model is fit, R^2 values are expected to decrease with this procedure. In addition, variables that may have been included in the baseline models because of the spatial structure would be not be included in the Cochrane–Orcutt models.

This procedure transforms the dependent and explanatory variables by using an estimate of the correlation between adjacent residuals

$$Y_t' = \beta_0' + \beta_1' X_t' + \mu_t \tag{2}$$

where

$$Y_t' = Y_t - \rho Y_{t-1} \qquad X_t' = X_t - \rho X_{t-1}$$

$$\beta_0' = \beta_0(1 - \rho) \qquad \beta_1' = \beta_1$$

and μ_t is the uncorrelated error term

$$e_t = \rho e_{t-1} + u_t \tag{3}$$

where e_t is the model residual associated with observation t, e_{t-1} is the model residual associated with the closest upstream observation, and u_t is the uncorrelated error term. The spatial correlation coefficient ρ is estimated by r

$$r = \frac{\sum_{t=2}^{n} e_{t-1} e_t}{\sum_{t=2}^{n} e_{t-1}^2} \tag{4}$$

The second approach (Stream ID) included individual stream identifiers as additional explanatory variables. In this expanded version of the base model (Eqn. 1), the suite of explanatory variables included indicator variables for the 40 individual streams in the data set in addition to the fish and macroinvertebrate variables. Stream identifiers were based on U.S. Geological Survey topographic maps (scale 1:250,000).

The last approach (Lag) included the closest upstream value of the dependent variable as one of the explanatory variables [7]. More complex models can incorporate the relationship between distance and the correlation between adjacent points; as the distance between adjacent observations increases, variable values are expected to become less similar [6,7,10]. However, exploratory plots (i.e., variograms [5]) indicated no clear structure, so we used the simplest approach and included a first-order lagged value, unweighted for the distance between observations

$$Y_t = \beta_0 + \alpha Y_{t-1} + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_t \tag{5}$$

where Y_t is the value of the dependent variable at location t, α is the first-order autoregressive coefficient, Y_{t-1} is the closest upstream value of Y, and the other variables are as defined in Equation 1. By incorporating spatial information more explicitly into the models, these Stream ID and Lag approaches would be expected to yield higher R^2 values.

Within each stressor factor and modeling approach, several models were fit by using a modified stepwise variable selection procedure that was designed to select models while controlling for correlation between the biological explanatory variables. Up to four seed biological variables were selected (based on their high correlation with the stressor factor) to initiate the modeling process. Standard stepwise variable selection procedures were then used; however, variables that were highly correlated (i.e., $p < 0.01$) with variables that were already in the model were not considered for addition [4]. This procedure resulted in up to four models for each modeling approach–stressor factor combination. Final models were selected based on testing performance (described below) and consistency of explanatory variables.

Table 1. Final models for stream corridor structure[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.21	*0.29*	CYPRINID	0.14
			OLIGO	−1.49
Cochrane–Orcutt	0.18	−0.068	CYPRINID	0.12
			OLIGO	−1.35
Stream ID[d]	0.53	−0.05	CYPRINID	0.149
	0.38[e]		OLIGO	−1.42
			SUNFISH	0.13
Lag	0.31	−0.055	CYPRINID	0.11
			OLIGO	−1.34

[a] Complete specifications for the models are provided in Appendices 2 and 3.
[b] D denotes the correlation between model residuals from adjacent stream locations. Correlations significant at $p = 0.05$ are shown in italic.
[c] Variables are defined in Appendix 1.
[d] Stream ID = Stream identifier.
[e] The R^2 value adjusted for the number of variables [5].

Table 3. Final models for TSS, Fe, and BOD[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.34	*0.47*	TOXTOL	−1.89
			HEADWATR	−1.05
			RELWT	−0.069
Cochrane–Orcutt	0.12	−0.04	TOXTOL	−1.18
			HEADWATR	−0.26
Stream ID	0.76	0.13	TOXTOL	−1.29
	0.69[d]		HEADWATR	−0.41
Lag	0.66	*−0.27*	HEADWATR	−0.30
			OTHDIP	−0.31
			RELWT	−0.023
			TOPCAR	0.06

[a] Complete specifications for the models are provided in Appendices 2 and 3; TSS = total suspended solids; Fe = iron; BOD = biochemical oxygen demand; Stream ID = Stream identifier.
[b] D denotes the correlation between model residuals from adjacent stream locations. Significant correlations are shown in italic.
[c] Variables are defined in Appendix 1.
[d] The R^2 value adjusted for the number of variables [5].

Model testing

The models were tested against data from a previous time period and against a random subset of the data. The first approach tested models that had been fit with all of the 179 observations available for years 1988 to 1994 (n =179) against data collected from 1980 to 1987. A limitation of the test data set was that only 28 samples representing seven streams had complete data for the variables of interest. When the additional constraints of the lagged variable approach were included (i.e., the observations in the test data set must be downstream of those in the training data set), the sample size decreased to 15. All of the instream chemistry variables in samples from 1980 to 1987 had lower or similar concentrations as the 1988 to 1994 data set except for Fe. Iron concentrations were substantially higher in the earlier time period. They ranged from 300 to 3,380 μg/L during 1980 to 1987 compared with 22 to 743 μg/L from 1988 to 1994. Before testing, the variables in the 1980 to 1987 data set were transformed to near-normality and regressed against drainage area by using the same relationships as used in the training data sets.

For the second testing approach, the 1988 to 1994 data set was randomly split 80:20, resulting in a data set of n = 143 for model building. Models were fit by using the procedures described above and then were tested against the 36 withheld samples.

The expected predictive performance of the models was evaluated by calculating the mean square prediction residuals (MSPR) with the mean square error (MSE) of the training model [11]. To aid in this comparison, the ratio of the MSPR and the MSE was calculated. If the MSPR is less than or equal to the MSE (i.e., ratios less than or equal to 1), the residuals in the test data set are comparable to those of the training data set. If the MSPR is greater than the MSE (i.e., ratios greater than 1), errors associated with application of the model would be larger than expected. The MSPR values substantially greater than the MSE indicate that the models would have poor predictive performance.

Table 2. Final models for siltation[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.10	*0.30*	GLYP	−1.05
			INSECT	0.011
Cochrane–Orcutt[b]	0.07	0.11	GLYP	−1.28
Stream ID[d]	0.46	−0.038	GLYP	−1.52
	0.30[e]		INSECT	0.009
Lag	0.18	−0.12	GLYP	−1.00

[a] Complete specifications for the models are provided in Appendices 2 and 3.
[b] D denotes the correlation between model residuals from adjacent stream locations. Significant correlations are shown in italic.
[c] Variables are defined in Appendix 1.
[d] Stream ID = Stream identifier.
[e] The R^2 value adjusted for the number of variables [5].

RESULTS

Significant models were found for all factors and all model approaches. Biological variables selected for each model are shown in Tables 1 through 6 and complete specification of all of the models is provided in Appendices 2 and 3. Cyprinids and oligochaetes were important variables predicting the quality of stream corridor structure (Table 1). Percent *Glyptotendipes* and insectivorous fish were important for predicting siltation

Table 4. Final models for COD and BOD[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.29	*0.44*	DEP	−1.22
			PIONEERP	0.18
			MAYFLY	−0.09
			TANY	−0.138
Cochrane–Orcutt	0.19	−0.04	DEP	−1.53
			PIONEERP	0.12
			TANY	−0.096
Stream ID	0.68	−0.030	DEP	−1.58
	0.58[d]		OTHDIP	0.26
Lag	0.52	*−0.20*	DEP	−1.75
			PIONEERP	0.089
			OTHDIP	0.30

[a] Complete specifications for the models are provided in Appendices 2 and 3; COD = chemical oxygen demand; BOD = biochemical oxygen demand; Stream ID = Stream identifier.
[b] D denotes the correlation between model residuals from adjacent stream locations. Significant correlations are shown in italic.
[c] Variables are defined in Appendix 1.
[d] The R^2 value adjusted for the number of variables [5].

Table 5. Final models for Zn and Pb[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.13	*0.46*	ALLINT	−0.06
			NUMTAXA	−0.03
Cochrane–Orcutt	0.09	0.054	ALLINT	−0.057
			NUMTAXA	−0.02
Stream ID	0.57	−0.008	ALLINT	−0.06
	0.44[d]		NUMTAXA	−0.02
Lag	0.29	−0.039	ALLINT	−0.047

[a] Complete specifications for the models are provided in Appendices 2 and 3; Stream ID = Stream identifier.
[b] D denotes the correlation between model residuals from adjacent stream locations. Significant correlations are shown in italics.
[c] Variables are defined in Appendix 1.
[d] The R^2 value adjusted for the number of variables [5].

(Table 2). Percent headwater fish species, toxic tolerant invertebrate taxa, and noninsect and dipteran invertebrate taxa were important variables in the TSS, Fe, and BOD models (Table 3). The proportions of mayflies, midges in the tribe Tanytarsini, and depositional insect species were important variables for the COD and BOD models (Table 4). Percent intolerant fish and invertebrate taxa richness were important for predicting Zn and Pb concentrations (Table 5). Finally, percent round-bodied suckers, carnivorous fish, numbers of fish, and shredding invertebrates were selected for the N and P models (Table 6).

Within each stressor factor, the explanatory variables selected varied across the different modeling options. However, when substitution occurred, the original and substituted variables were always strongly correlated with each other. For example, the number of intolerant species was replaced by the percent of round-bodied suckers in the Stream ID model for N and P factor. For the Cochrane–Orcutt procedure, variables that had high correlations between adjacent points tended to be dropped from the models. For example, the percent of shredders did not meet the criteria for inclusion in the N and P model after it was corrected for spatial dependence. When the same explanatory variable appeared in several model approaches for the same stressor factor, the directionality of the parameter estimate remained consistent.

Table 6. Final models for NO_x and P[a]

Modeling approach	R^2	D[b]	Variables[c]	Parameter estimates
Base	0.13	*0.45*	TOPCARN	−0.125
			RDSUCKPC	−0.098
			RELNO	−0.39
			SHRED	0.247
Cochrane–Orcutt	0.11	0.003	TOPCARN	−0.128
			RDSUCKPC	−0.080
			RELNO	−0.48
Stream ID	0.56	0.0039	TOPCARN	−0.16
	0.42[d]		INTOLS	−0.11
Lag	0.36	0.0005	TOPCARN	−0.09
			RDSUCKPC	−0.057
			RELNO	−0.35

[a] Complete specifications for the models are provided in Appendices 2 and 3; NO_x = nitrate and nitrite; P = phosphorus; Stream ID = Stream identifier.
[b] D denotes the correlation between adjacent model residuals. Significant correlations are shown in italic.
[c] Variables are defined in Appendix 1.
[d] The R^2 value adjusted for the number of variables [5].

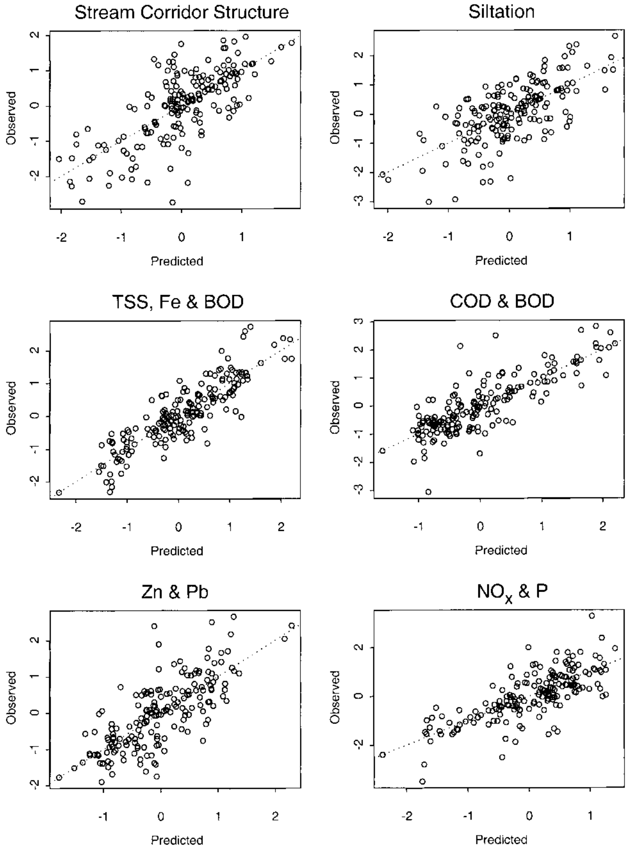

Fig. 1. Plots of observed versus predicted values for the six stressor factors with the stream identifiers modeling approach. TSS = total suspended solids; BOD = biochemical oxygen demand; COD = chemical oxygen demand; NO_x = nitrate and nitrite.

The R^2 values were the highest for the Stream ID models, ranging from 0.30 for the siltation factor to 0.69 for the TSS, Fe, and BOD factor. Plots of predicted versus observed values for these models are shown in Figure 1. Standard regression diagnostic plots were produced for each of the final models (residuals vs fitted values, quantiles of residuals vs standard normal distribution, Cook's distance) [5,11]. They indicated that no issues (e.g., overly influential observations) required mitigation.

Compared with the base approach, the Cochrane–Orcutt procedure generally decreased the R^2 values, whereas the Stream ID code and the Lag modeling approaches increased the R^2 of the models. The Cochrane–Orcutt, Stream ID, and Lag modeling approaches markedly decreased the correlation between adjacent residuals (ρ); the correlations associated with most models were not significant at $p = 0.05$.

The application of the $n = 179$ training set to the 28 samples collected from earlier years showed mixed success. Almost one half (11 of 24) of the model tests had MSPR:MSE ratios greater than 1.5. The MSPR:MSE ratios (Table 7) were near or less than 1 for the stream corridor structure factor but were uniformly high for the siltation factor. At least one model under each approach (except for Stream ID) had an MSPR:MSE ratio near 1 for the TSS, Fe, and BOD factor. Only the base model approach and the Cochrane–Orcutt procedure had MSPR:MSE ratios that were near 1 for the COD and BOD factor. Only the base model had MSPR:MSE ratios that were near 1 for the Zn and Pb factor and the NO_x and P factor.

Models constructed with the $n = 143$ data set (not shown)

Table 7. The MSPR: MSE ratios of test data (*n* 28 from 1980–1987) and final models (n = 179)[a]

Factor	Base	Cochrane–Orcutt	Stream ID	Lag
Stream corridor structure	0.67	0.64	0.88	0.63
Siltation	1.48	2.02	2.10	1.96
TSS, Fe, and BOD	0.80	1.05	2.40	1.03
COD and BOD	1.01	0.92	1.74	1.33
Zn and Pb	0.95	2.08	2.76	2.70
NO_x and P	0.93	1.71	2.82	1.90

[a] MSPR = means square prediction residuals; MSE = mean square error; Stream ID = Stream identifier; TSS = total suspended solids; BOD = biochemical oxygen demand; COD = Chemical oxygen demand; NO_x = nitrate and nitrite.

were similar to those constructed with the full data set, with some substitution of strongly correlated variables. The models generally performed well with the randomly selected subset of 36 observations; 20 of the 24 model tests had MSPR:MSE ratios less than 1.5 (Table 8). The Stream ID models performed very well for the TSS, Fe, and BOD factor and COD and BOD factors. The Lag models performed well across all of the stressor factors.

DISCUSSION

This paper presents a series of models that use biological assessment data to predict the degree of six types of anthropogenic stress. Significant models were found for all six stressor factors explored in this study. The overall consistency in explanatory variables and paramater estimates across the different modeling approaches increases the confidence that the models are biologically meaningful.

The models with the highest R^2 values included either the Stream ID or the lagged dependent variable. The great improvement in R^2 values seen in the last two modeling options has several interpretations. One is that not all of the important variables were included in the base models. Stream ID may be acting as a surrogate for variables that might represent the initial condition of the stream community, other stressor variables, or attributes of the stream that might mitigate response to stress. The lagged variable has a more physical, source-related interpretation. The level of stress at any site logically is affected by the level of stress at upstream sites, which reflects underlying fate and transport processes within the stream.

The parameter estimates are difficult to interpret quantitatively because of the data transformations and the regression against drainage area. However, in terms of directionality, most variables and parameter estimates are consistent with biological expectation.

Sites with high scores for quality of stream corridor structure (Table 1) were associated with more minnows and fewer oligochaetes. The increase in minnows may be related to the existence of refugia at sites scoring high for stream corridor structure. The increased number of oligochaetes at sites scoring low for stream corridor structure generally is consistent with the classification of oligochaetes as tolerant organisms.

Sites with high siltation were associated with a higher percent of *Glyptotendipes* and fewer insectivorous fish (Table 2). High proportions of *Glyptotendipes* have been associated with agricultural nonpoint sources and with conventional municipal waste treatment plants, perhaps because of the nutrients contributed by both of these sources [12]. The increase in *Glyptotendipes* may provide evidence that the primary source of sediments in the Eastern Corn Belt Plains ecoregion is from agricultural soils that contain high concentrations of nutrients. The relationships between siltation and the biological variables were quite weak. A potential explanation for this is that the use of the Hester–Dendy samplers mitigated the impacts of siltation on the macroinvertebrates by providing hard surface habitat. Still, the finding is inconsistent with another study in the same region that found that the substrate metric was an important predictor variable of index of biotic integrity (fish community) scores [13]. However, the final explanatory variables used in that study did not include drainage area or a surrogate (e.g., stream order), which may explain the inconsistency.

The models for the TSS, Fe, and BOD factor (Table 3) had the highest R^2 values, but presented difficulties in interpretation. The proportion of headwater species decreased with higher concentrations of TSS, Fe, and BOD. Headwater species are considered to be sensitive but are also strongly influenced by stream gradient, and their inclusion in the model may indicate a residual effect of this covariate. The number and weight of fish decreased at higher concentrations. The percent of toxic-tolerant invertebrate taxa and noninsect and dipteran invertebrate taxa also decreased with increasing TSS, Fe, and BOD. This finding is inconsistent with initial expectations, because these organisms are considered to be tolerant of chemical stress. High concentrations of suspended solids may reduce the bioavailability of toxic substances in the water column. The increase in suspended solids actually may reduce the effective toxicity of chemicals in the water column, yielding these counterintuitive results.

Increasing values for COD and BOD (Table 4) were associated with decreasing proportions of mayflies, midges in the tribe Tanytarsini, and depositional insect species. Mayflies and midges in the tribe Tanytarsini appear frequently in the literature as sensitive indicators of stressors such as nutrients, sources such as sewer overflows and industrial outfalls, and land uses such as agriculture and urbanization [14–18]. The percent of pioneering species, considered to be tolerant of stress, increased with increasing concentrations of COD and BOD.

Percent intolerant fish decreased with increasing Zn and Pb concentrations (Table 5). The percent of round-bodied suckers and intolerant fish species also decreased with increasing concentrations of NO_x and P (Table 6). Although both of these results are consistent with expectations in general terms, seeing such similar explanatory variables indicates that the effects of these two stressor factors may be difficult to differentiate in the Eastern Corn Belt Plains ecoregion. The percent of shredders increased with higher concentrations of NO_x and P. If

Table 8. The MSPR: MSE ratios of test data (n = 36, randomly selected) and final models (n = 143)[a]

Factor	Base	Cochrane–Orcutt	Stream ID	Lag
Stream corridor structure	1.06	1.27	2.23	1.09
Siltation	0.82	0.77	1.62	0.77
TSS, Fe, and BOD	0.78	1.10	0.49	1.00
COD and BOD	1.10	0.62	1.10	1.38
Zn and Pb	1.05	1.23	1.63	1.10
NO_x and P	1.16	1.08	1.87	1.19

[a] MSPR = mean square prediction residuals; MSE = mean square error; Stream ID = Stream identifier; TSS = total suspended solids; BOD = biochemical oxygen demand; COD = chemical oxygen demand; NO_x = nitrate and nitrite.

these nutrients increased plant growth, the detritus that shredders feed on may also increase, or become more nutritive through increased microbial colonization.

The results of this analysis were consistent with the discriminant function results described previously [1]. In that study, sites were grouped into low-, medium-, and high-stress categories based on quartiles of each stressor factor distribution. In most cases, the variables selected in this study were among the most strongly correlated with the discriminant function for the same stressor factor. In a few cases, additional variables proved important in the models. These included percent top carnivores in the TSS, Fe, and BOD factor models; percent mayflies in the COD and BOD factor models; and the number of fish and percent shredders in the NO_x and P factor models. Some of these differences can be attributed to the clustering of sites into groups that was necessary for the discriminant analysis. The regression approach has the advantage of treating the stressor factor scores as continuous.

The strong testing results found by randomly partitioning the 1988 to 1994 data set indicates that these models will have predictive value for estimating the degree of stress at locations where biological samples are available, but stream chemistry or habitat variables are not, within this time period. In cases where predictions are desired in streams that have at least one existing sample, the use of the Stream ID or Lag models would provide more accurate predictions.

In contrast, the mixed testing results and the generally high MSPRs for the Lag and Stream ID models when tested against the observations collected from 1980 to 1987 indicate that these models will have little predictive value for additional observations from that time period. The particularly high MSPRs for the Lag and Stream ID models indicate that the spatial structure of the data changed from the earlier time period to the later. This conclusion is supported by an analysis of the historical trends in one of the streams—the Big Darby Creek [19]. In that study, a higher degree of spatial correlation was seen in the time periods of 1986 to 1993, which would roughly correspond to the period of time of the training data set, as compared to an earlier time period of 1979 to 1981. Differences in spatial structure may be attributable to continued refinement of the biological sampling methods during the early 1980s (M. Smith, personal communication). However, management actions, including the removal of low-head dams and the institution of additional wastewater treatment, also occurred in the mid- to late 1980s. These changes likely also changed the spatial structure of the data by increasing connectivity in the streams.

Finally, the modeling results provide insight into whether biological communities respond in distinctive ways to different types of stress. In this study, very different biological variables and parameter estimates best explained the variability for four of the stressor factors: stream corridor structure; siltation; TSS, Fe, and BOD; and COD and BOD. This indicates that the biological communities may be responding differently to these types of stress. The model variables and parameter estimates that fit for the Zn and Pb factor and the NO_x and P factor were very similar, with both relying on intolerant fish species. This indicates that these two types of stress may be difficult to distinguish with these models.

The models produced in this effort are product of the specific stressors, processes, and biological communities present in the Eastern Corn Belt Plains ecoregion of Ohio. The same models would be unlikely to accurately predict the degrees of different types of stress when applied to other regions. However, the relationships seen should provide relevant insights into the general patterns of biological responses that we can expect in response to these stressors. Finally, the models provide additional evidence that biological communities can serve as useful indicators of the types anthropogenic stress that are impacting aquatic systems.

Acknowledgement—We gratefully acknowledge the assistance of Dennis Mishne, Ed Rankin, Jeff DeShon, Steve Gordon, Sarada Majumder, Florence Fulk, and Karen Blocksom, and the comments of Michael Troyer, Rick Racine, and two anonymous reviewers.

REFERENCES

1. Norton SB, Cormier SM, Smith M, Jones RC. 2000. Can biological assessments discriminate among types of stress? A case study from the Eastern Corn Belt Plains ecoregion. *Environ Toxicol Chem* 19:1113–1119.
2. Norton SB. 1999. Using biological monitoring data to distinguish among types of stress in the Eastern Corn Belt Plains ecoregion. PhD thesis. George Mason University, Environmental Science and Public Policy, Fairfax, VA, USA.
3. Cleveland WS. 1993. *Visualizing Data.* Hobart Press, Summit, NJ, USA.
4. SAS Institute. 1990. *SAS® User's Guide: Statistics, Ver 6.* Cary, NC, USA.
5. Statistical Sciences. 1993. *S-Plus for Windows User's Manual:* Ver 3. 1. Statistical Sciences, Seattle, WA, USA.
6. Diggle PJ. 1990. *Time Series: A Biostatistical Introduction.* Oxford University Press, New York, NY, USA.
7. Anselin L. 1988. *Spatial Econometrics: Methods and Models.* Kluwer, Dordrecht, The Netherlands.
8. Edbon D. 1985. *Statistics in Geography.* Blackwell, Cambridge, MA, USA.
9. Thomann RV, Mueller JA. 1987. *Principles of Surface Water Modeling and Control.* Harper Collins, New York, NY, USA.
10. Cressie N. 1991. *Statistics for Spatial Data.* John Wiley, New York, NY, USA.
11. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. 1996. *Applied Linear Statistical Models.* Irwin, Chicago, IL, USA.
12. Yoder CO, Rankin ET. 1995. Biological response signatures and the area of degradation value: New tools for interpreting multimetric data. In Davis WS, Simon TP, eds, *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making.* Lewis, Boca Raton, FL, USA, pp 263–286.
13. Majumder S. 1998. A spatial empirical analysis of stressor–response relationships for prospective ecological risk assessment in the Eastern Cornbelt Plains of Ohio. PhD thesis. Ohio State University, Columbus, OH, USA.
14. Lenat DR, Crawford JK. 1994. Effects of land use on water quality and aquatic biota of three North Carolina piedmont streams. *Hydrobiologia* 294:185–199.
15. Quinn JM, Hickey CW. 1990. Magnitude of effects of substrate particle size, recent flooding, and catchment development on benthic invertebrates in 88 New Zealand rivers. *N Z J Mar Freshwater Res* 24:411–427.
16. Townsend CR, Arbuckle CJ, Crowl TA, Scarsbrook MR. 1997. The relationship between land use and physicochemistry, food resources and macroinvertebrate communities in tributaries of the Taieri River, New Zealand: A hierarchically scaled approach. *Freshwater Biol* 37:177–191.
17. Jones RC, Clark CC. 1987. Impact of watershed urbanization on stream insect communities. *Water Resour Bull* 23:1047–1055.
18. Yoder CO, Rankin ET. 1998. The role of biological indicators in a state water quality management process. *Environ Monit Assess* 51:61–88.
19. Schubauer-Berigan MK, Smith M, Hopkins J, Cormier SM. 1999. Using historical biological data to evaluate status and trends in the Big Darby Creek (Ohio) watershed. *Environ Toxicol Chem* 20:1097–1081.
20. Ohio Environmental Protection Agency. 1989. *Biological Criteria for the Protection of Aquatic Life:* Vol III: Standardized Biological and Field Sampling and Laboratory Methods for Assessing Fish and Macroinvertebrate Communities. Procedure WQPA-SWS-3. Division of Water Quality Planning and Assessment, Ecological Assessment Section, Columbus, OH, USA.

APPENDIX 1. Biological variables used in models[a]

Variable	Description	Transformation used	Regressed against drainage area?
ALLINT	Number of sensitive fish sp. (i.e., ''I'' and ''M'' in Ohio Environmental Protection Agency [Ohio EPA] species file)	None	Yes: quadratic
CYPRINID	Number of minnow species	None	Yes: Quadratic
HEADWATR	Number of headwater fish species in a sample	Log	Yes: linear
INSECT	Percent insectivorous fish	None	Yes: quadratic
INTOLS	Number of intolerant fish sp. (i.e., ''I'' in Ohio EPA species file)	None	Yes: quadratic
PIONEERP	Percent individuals of pioneering fish species in a sample	Square root	Yes: linear
RDSUCKPC	Percent round-bodied suckers in a sample	Square root	Yes: linear
RELNO	Number of fish per unit distance	Log	Yes: linear
RELWT	Weight of fish per unit distance	Square root	Yes: linear
SUNFISH	Number of sunfish species	None	Yes: quadratic
TOPCARN	Percent carnivorous fish in a sample	Square root	Yes: linear
DEP	Percent of total number of invertebrates that are depositional taxa	$x^{0.25}$	Yes: linear
GLYP	Percent of total number of invertebrates that are *Glyptotendipes*	$x^{0.25}$	Yes: linear
MAYFLY	Percent of total number of invertebrates that are mayfly taxa	Square root	Yes: quadratic
NUMTAXA	Total number of invertebrate taxa	None	No
OLIGO	Percent of the total number of invertebrates that are oligochaetes	$x^{0.25}$	Yes: quadratic
OTHDIP	Percent of total number of invertebrates that are dipterans and noninsects	Log	Yes: linear
SHRED	Percent of total number of invertebrates that are shredding insect taxa	$x^{0.25}$	No
TANY	Percent of total number of invertebrates that are midges in the tribe Tanytarsini	Square root	Yes: linear
TOXTOL	Percent of the total number of invertebrates that are toxic tolerant	$x^{0.25}$	Yes: linear

[a] Fish were sampled by electrofishing a standard length of stream, and invertebrates were sampled with Hester–Dendy artificial substrate samplers [20].

APPENDIX 2. Parameter estimates for final models: linear regression, Cochrane–Orcutt, and lag variable models[a]

	Linear		Cochrane–Orcutt		Lag	
Factor	Variable	Estimate	Variable	Estimate	Variable	Estimate
Stream corridor structure	Intercept	−0.02645255	Intercept	0.01999269	Intercept	0.02237040
	CYPRINID	0.13964855	CYPRINID	0.12275532	Lag	0.26907438
	OLIGO	−1.49003922	OLIGO	−1.35352723	CYPRINID	0.11159972
					OLIGO	−1.33683716
Siltation	Intercept	0.03597585	Intercept	−0.00337758	Intercept	−0.00127199
	GLYP	−1.04670787	GLYP	−1.27996383	Lag	0.31097454
	INSECT	0.01135981			GLYP	−1.00714717
TSS, Fe, and BOD	Intercept	0.09890619	Intercept	0.01394814	Intercept	0.01816708
	TOXTOL	−1.89898569	TOXTOL	−1.18575477	Lag	0.59953495
	HEADWATR	−1.04728530	HEADWATR	−0.26274274	HEADWATR	−0.30390340
	RELWT	−0.06888591			OTHDIP	−0.03121970
					RELWT	−0.02352797
					TOPCAR	0.05957139
COD and BOD	Intercept	0.07216969	Intercept	−0.01540381	Intercept	−0.03241357
	DEP	−1.22068150	DEP	−1.53665446	Lag	0.50199922
	PIONEERP	0.18158845	PIONEERP	0.12083238	DEP	−1.75314809
	MAYFLY	−0.09324662	TANY	−0.09621573	PIONEERP	0.08922815
	TANY	−0.13841174			OTHDIP	0.29757691
Zn and Pb	Intercept	0.96872162	Intercept	0.42454920	Intercept	0.01735244
	ALLINT	−0.05394964	ALLINT	−0.05711383	Lag ALLINT	0.46533381
	NUMTAXA	−0.02599705	NUMTAXA	−0.01761057		−0.04703134
NO_x and P	Intercept	−0.14333429	Intercept	0.05159821	Intercept	0.06733056
	TOPCARN	−0.12507637	TOPCARN	−0.12849238	Lag	0.44340646
	RDSUCKPC	−0.09813103	RDSUCKPC	−0.07958748	TOPCARN	−0.09052676
	RELNO	−0.38632408	RELNO	−0.47811403	RDSUCKPC	−0.05662102
	SHRED	0.24740824			RELNO	−0.34991299

[a] Variables are defined in Appendix 1; TSS = total suspended solids; BOD = biochemical oxygen demand, COD = chemical oxygen demand; NO_x = nitrate and nitrite.

APPENDIX 3. Parameter estimates for final models: Stream identifier (Stream ID) models[a]

	Parameter estimates					
Variable	Stream corridor structure	Siltation	TSS, Fe, and BOD	COD and BOD	Zn and Pb	NO_x and P
Intercept	−0.13832	0.73621	−0.47070	−0.17244	−1.11789	−2.38698
R01_001	0.23258	−2.87121	0.07938	−0.58324	0.49212	1.98291
R02_001	0.01478	−1.40565	0.26025	0.26896	2.54876	2.95151
R02_069	−0.45842	−0.24102	−0.51276	−0.33583	1.41227	2.21760
R02_100	0.40057	−0.63096	0.89464	−0.46736	2.42228	1.41381
R02_109	0.21284	−0.57209	0.43742	1.78086	3.51575	3.15929
R02_138	0.18139	−0.80117	1.27789	0.34144	0.52495	2.91421
R02_200	0.30138	−0.18803	0.64115	−0.10655	1.69185	2.95129
R02_204	0.00533	−1.39285	−0.01615	0.07644	1.18871	2.98486
R02_207	0.72438	−0.94588	0.36015	2.04947	2.23704	1.91627
R02_210	0.96912	0.68045	0.12643	−0.25846	2.99707	2.54316
R02_211	0.82854	−0.56214	−0.17033	−0.11336	0.99936	3.17029
R02_245	0.22640	−0.59981	0.26492	0.16343	0.11115	2.87123
R02_400	0.71712	−1.09356	0.55552	0.38431	1.54499	2.21479
R02_500	−0.61610	−0.51990	0.87093	−0.78360	1.76447	3.56693
R04_100	−0.53285	−0.32446	1.50590	1.00706	1.00263	1.22028
R04_160	−0.69266	0.39174	0.51196	0.30670	2.34784	3.24677
R04_168	0.18251	−0.13204	0.03550	0.75101	2.19421	0.66129
R04_200	0.93667	0.21687	0.36231	1.34094	1.56977	2.20911
R04_221	0.87734	0.30182	0.11114	1.40664	1.59064	2.90545
R04_500	0.60381	−0.19250	1.49075	1.97845	1.53047	1.83830
R05_001	0.31761	−1.35300	1.23285	−0.30789	2.42373	1.88610
R11_001	0.58400	−0.10206	0.23014	−0.11212	1.82671	2.83103
R11_040	0.96371	0.23852	−0.49322	0.88966	2.74914	3.50999
R14_001	−0.50360	−1.44231	0.65607	1.40150	1.02993	1.14505
R14_043	0.06205	−1.05509	−0.45680	−0.19528	1.41889	2.93159
R14_100	0.49696	−0.89857	−0.50141	−0.65504	1.53569	2.45251
R14_110	−1.19484	−0.82392	−0.55843	0.26939	0.85883	1.92109
R14_120	−0.90821	−1.77609	−1.35892	−0.01798	2.31634	1.88917
R14_130	−0.25000	−0.85427	−0.32586	−0.39862	1.80362	2.71530
R14_139	−0.37439	−0.50203	−1.22770	−1.70213	1.29954	2.60600
R14_200	−0.14308	−0.19281	1.43522	0.05287	1.65923	2.77636
R14_208	−0.65938	0.76332	1.30825	0.37230	2.00498	3.00604
R14_220	0.67062	−0.59810	1.41493	−0.14842	1.80288	3.17934
R14_226	−0.72319	−0.08962	2.51133	−0.87223	2.26296	2.29024
R14_227	0.76114	−1.11363	1.24876	−0.38875	1.01038	3.13625
R14_235	−1.56217	−0.33275	2.38426	0.55406	2.88834	3.24118
R14_400	1.23322	0.81730	0.62501	1.19479	2.92226	1.68723
R14_410	0.48859	−0.41887	−0.29473	0.04476	2.84301	2.78308
R14_600	−0.27234	−2.20998	1.01598	1.18842	1.12120	2.89603
R14_804	−1.49138	−0.36390	1.27986	2.57288	1.62348	1,57901
CYPRINID	0.14653	—	—	—	—	—
OLIGO	−1.44217	—	—	—	—	—
SUNFISH	0.12931	—	—	—	—	—
GLYP	—	−1.53833	—	—	—	—
INSECT	—	0.00892	—	—	—	—
TOXTOL	—	—	−1.26984	—	—	—
HEADWATR	—	—	−0.37400	—	—	—
DEP	—	—	—	−1.46668	—	—
OTHDIP	—	—	—	0.26632	—	—
ALLINT	—	—	—	—	−0.05523	—
NUMTAXA	—	—	—	—	−0.01726	—
TOPCARN	—	—	—	—	—	−0.16093
INTOLS	—	—	—	—	—	−0.11649

[a] Variables beginning with R refer to Stream ID. Other variables are defined in Appendix 1. TSS = total suspended solids; BOD = biochemical oxygen demand; COD = chemical oxygen demand; NO_x = nitrate and nitrite.