Development of a Predicting Model for Biotic Integrity Based on Variable Selection with Self-Organizing Maps(SOM), Polynomial Canonical Correspondence Analysis (PCCA) and Quadratic Regressions

TECHNICAL REPORT NO. 12

David Bedoya, PhD Candidate

Center for Urban and Environmental Studies
Northeastern University
Boston,MA

Project sponsored by the Grant No. R83-0885-010 to Northeastern University from the
US EPA/NSF/USDA STAR Watershed Program

Vladimir Novotny, Ph.D., P.E.
Primary Investigator
Center for Urban Environmental Studies
Northeastern University, Boston, MA

Bernice Smith
EPA Project Officer

Iris Goodman
EPA Program Director

June 2007

Abstract

A model to predict biotic integrity in the states of Minnesota, Ohio and Maryland was developed. The model was based in three different steps. The first one consisted of selecting the most relevant environmental variables. The selection was performed in two different ways: clustering followed by analysis of the most discriminant metrics among the clusters found with multiple range tests and polynomial canonical correspondence analysis, which selects those metrics that have greater effect over the biotic community based on regression, eigenvalue decomposition and projection of the sites over the explanatory variables in the canonical axes. Once the most selective variables were identified, a polynomial regression was performed trying to approximate the index of biotic integrity (fish or benthic) in that site. The observed values were then separated in four different data range bins. The calculated values were then located into one of those bins and the number of correct bin-predictions were recorded.

The model proved to work well especially in the states of Minnesota and Ohio. In the case of Maryland the model didn't work as well probably due to problems with biased habitat parameters belonging to the old habitat quality index and potentially inaccurate parameters in the new one. Also, the model seemed to work better with actual physical parameters measurements than with habitat quality scores. Actual measurements were available only in the Minnesota's database. Even though not much difference existed between the two different ways of selecting metrics, clustering seemed to get more accurate regression parameters but canonical analysis seemed to work slightly better in the binning method. Further improvement of the model needs to be done by removing outliers and by taking sub-datasets that have the same number of sites in each of the four bins to avoid better predictions for bins with larger number of sites in it.

Acknowledgements

The research contained in this report was sponsored by the U.S. Environmental Protection Agency/National Science Foundation/U.S. Department of Agriculture STAR Watershed Program by a Grant (No. R83-0885-010) to Northeastern University. The authors greatly appreciate this support. The findings and conclusions contained in this report are those of the authors and not of the funding agencies, nor the STAR program. The authors would like to express their thanks to all agencies that kindly provided their data, including the Maryland's Department of Natural Resources, Maryland Biological Stream Survey, Ohio Environmental Protection Agency, Minnesota Pollution Control Agency, and United States Geological Survey.

Table of contents

1. Introduction

In the stressors-risk-endpoints-propagation model, water body integrity had three dimensions: chemical, physical/habitat and biotic integrity (Novotny et al. 2005). Five different human-induced stresses on water resources are known for being responsible of stream integrity degradation: water quality, habitat structure, flow regime, energy sources, and biotic interactions (Karr and Kerans, 1991).

Human intervention within the drainage area or adjacent regions of a stream will have an effect on instream habitat, water quality and finally on biological community. These changes can take place in the watershed area (i.e. increase of the urban or impervious area, reduction of the riparian buffer width) or in the stream channel itself (i.e. channelization or impoundments). One of the most challenging issues in stream biological integrity assessments is to reflect the effect of the different stressors over the biological community. Water quality is easily measurable and it's been widely studied. Point sources are easily to identify and control with the NPDES permits as part of a Total Maximum Daily Load (TMDL) process. Non-point sources, however, are more challenging because they are usually related to land use and runoff from wet weather events. The runoff from agricultural or paved areas will have a deep impact in the stream if preventive measures are not taken (riparian strips or best management practices).

Instream habitat quality is usually measured with a multi-metric index. In the U.S., these indices are usually state-based. Examples are the Qualitative Habitat Evaluation Indices (QHEI) in Ohio or the Physical Habitat Indices (PHI) in Maryland (Rankin, 1989, Hall Jr. et al., 1999, Paul et al., 2003). Even though a great variety of stream habitat indices and sampling methodologies exist (Kauffman et al., 1999, Lazorchak et al., 2000, Barbour et al.,1999), efforts have been made in unifying criteria and simplifying habitat quality evaluation with methodologies such as the *Rapid Bioassessment Protocols for Use in Wadeable Streams and Rivers* by Barbour et al. (1999).

The habitat indices are based mostly based on qualitative evaluation of different physical parameters, whose changes have an impact on stream fauna. The parameters are scored by experts on a scale depending on their degree of variation from reference sites. The scores for the different parameters (called metrics) are then summed and scaled (in the case of Maryland) to an understandable range for different strata. In the case of Minnesota and Ohio's QHEI and Maryland's PHI, the indices ranges from 0 (very poor) to 100 (excellent). This *multimetric* approach is used in more than 85% of water quality programs in the U.S.A (Southerland and Stribling, 1995).

Another approach for habitat evaluation is using predictive models, mainly in other countries such as the United Kingdom (RIVPACS), Australia (AUSRIVAS), or Canada (BEAST), (Norris and Hawkins, 2000). Predictive models quantify river health as to the degree to which a site would support the biota that would be expected if no human intervention existed. In this approach an empirical model is created that considers environmental variables of sites where human degradation is unlikely and relates these variables to the probability of a species occurring therein. These environmental variables are then used to match test sites with reference sites so that site-specific predictions of expected taxonomic composition can be made (Norris and

Hawkins, 2000). Some discussion has arisen on the accuracy and reliability of both approaches (Karr and Chu, 2000, Norris and Hawkins, 2000).

Biotic integrity is also measured in the U.S. with a multimetric approach. The Index of Biotic Integrity (IBI) is composed of several metrics that should reflect the impact of the stressors on the community being sampled. The IBI is usually performed with either fish or macroinvertebrate communities. The fish IBI in the U.S. is normally based in the index developed by Karr (1986). This index is comprised of twelve metrics divided in three categories: species richness and composition, trophic composition and fish abundance and condition. It consists of fish sampling and scoring for each metric. The scoring is based on a scale in which the highest scores correspond to sites that resemble reference sites and viceversa. Many states have developed their own fish IBI (Ohio EPA, 1987, Niemela and Feist,2000, Niemela and Feist, 2002, Roth et al., 2000). Also a myriad of benthic community indices exist. Some examples are the Hilsenhoff index (Hilsenhoff, 1987), the ICI or Ohio's Invertebrate Community Index (Ohio EPA, 1987), the Benthic Index of Biotic Integrity (BIBI) in Maryland (Stribling et al., 1998), or the Macroinvertebrate Index of biological Integrity (MIBI) in Minnesota (Chirart, 2003, Genet and Chirart, 2004). The advantages of measuring macroinvertebrates instead of fish are that they are relatively immobile, easy to collect at low cost, they occupy all stream habitats and are quick to react to environmental change (Ohio EPA, 1987, Mason, 1991).

2. Goals and objectives

Stream's biotic integrity is the result of a combination of factors that may include environmental parameters of different nature. Chemical water quality, instream habitat suitability, immediate types of land use along the stream or land use distribution, or hydrologic regime changes are all easily measurable parameters which are part of an ecological system in which any change in one of them will affect the ultimate water body integrity (Karr et al., 1986, Novotny, 2003). Even though it is known that they are all intertwined, the cause-effect relationships are not well understood and they are probably too difficult to be fully comprehended. The identification of the main explanatory environmental variables affecting biotic integrity in each site or homogeneous group is paramount. However, they are not necessarily the same in different sites with similar biotic integrity. Methodologies able to identify these parameters with no a priori assumptions is key. SOM, PCCA and multivariate regressions are methods that have these qualities and, therefore, were used in our analysis.

In our research we tried to predict biotic integrity based on non-biotic parameters with no a-priori assumptions. The research will help us understand which parameters are the most important for benthic and fish communities in the states of Ohio, Maryland and Minnesota. On the other hand, the use of SOM will help us identify different functional units (or clusters) with similar physical and chemical characteristics. Subsequently, our research focused on the identification of the main environmental parameters responsible for the biotic integrity quality in the functional units using Multiple range Tests (MRT). Also PCCA was used in order to identify the principal variables affecting fish in the whole state. Finally, obtaining predicting equations based on non-linear regressions to predict biotic integrity with the selected metrics was the last objective in our research.

The findings of the present paper should enable watershed managers predict or approximate some site's biotic integrity based on non-biotic parameters. Also, it would allow them to study the effect that a change in some of the environmental variables would have on biotic integrity, since its prediction would be modified by either a change of functional unit and therefore regression equation or just by entering a different parameter in the same regression equation in the same functional unit

3. Data description

Environmental databases were obtained from different agencies, Ohio EPA, Maryland Biological Stream Survey (MBSS), and Minnesota Pollution Control Agency (MNPCA). The three databases had records for habitat and water quality parameters, drainage area and/or adjacent cover land use and fish counts and biotic indices. Each state had different parameters measured and the habitat and biotic indices were also different. Descriptions of the habitat and biotic indices for each state can be found in available literature (Rankin, 1989, Ohio EPA 1987 and 1989, USEPA, 1998, Paul et al., 2003, MBSS, 1999, Niemela and Feist, 2000, and 2002, Simonson et al., 1994, Southerland et al., 2005) A more detailed description of the data included in each state's database are explained in Table 3-1, Table 3-2, and Table 3-3.

In Ohio 1,848 sites with data were available and the observations ranged from 1995 to 2000. 1,328 sites had observations for drainage area and only 429 sites had records for every field. In the case of Maryland the data ranged from 1995 to 1997 and had 905 sites with full records among the three strata (coastal, piedmont and highland areas). Minnesota had 1,134 observations with habitat scores, stream morphology or biotic quality. However only in 163 cases full records for habitat scores, water quality values and biotic integrity scores (fish IBI) were available. Also in Minnesota, 88 sites had observations for actual stream physical data measurements such as percentage of embeddedness.

3.1. Ohio's database

TYPE OF DATA	NAME	DESCRIPTION	UNITS
CHEMICAL PARAMETERS	TEMPERATURE	Water temperature	Degrees centigrade
	CONDUCTIVITY	Water conductivity	
	DO	Dissolved Oxygen	mg/L
	BOD	Biological Oxygen Demand	mg/L
	PH	Water pH	Standard units
	TSS	Total Suspended Solids	mg/L
	AMMONIA	Ammonia in water	mg/L
	NITRITE	Nitrite in water	mg/L
	TKN	Total Kjeldahl Nitrogen	mg/L
	NITRATE	Nitrate in water	mg/L
	PHOSPHORUS	Total Phosphorus in water	mg/L
	HARDNESS	Hardness in water	ppm
	CALCIUM	Dissolved calcium	mg/L
	MAGNESIUM	Dissolved Magnesium	mg/L
	CHLORIDE	Dissolved Chloride	mg/L
	SULFATE	Dissolved Sulfate	mg/L
	AS	Dissolved Arsenic	mg/L
	CD	Dissolved Cadmium	mg/L
	CU	Dissolved Copper	mg/L
	IRON	Dissolved Iron	mg/L
	PB	Dissolved Lead	mg/L
	ZN	Dissolved Zinc	mg/L
PHYSICAL/HABITAT PARAMETERS	SUBSTRATE	Substrate quality and type	Score from 0 to 20
	EMBEDDED	Degree to which the parent material is covered by fine sediment	Scale from 0 to 4
	COVER	Amount and type of stream vegetal cover	Score from 0 to 20
	CHANNEL	Quality of the stream with regard to creation and stability of macrohabitat	Score from 0 to 20
	RIPARIAN	Riparian zone width and type of vegetation and bank erosion	Score from 0 to 10
	POOL	Maximum depth of pool and type and morphology	Score form 0 to 12
	RIFFLE	Riffle depth, stability and embeddedness	Score from 0 to 8
	GRADIENT_S	Elevation drop through the sampling area	Score from 0 to 10
	PER_AG	Percentage of agriculture in beyond buffer area	0,25,50,75 or 100%
	PER_FORWET	Percent of forest and/or wetlands beyond buffer area	0,25,50,75 or 100%
	PER_URBDEV	Percentage of urban/developed beyond buffer area	0,25,50,75 or 100%
	AREA	Drainage area of the site	Square miles

Table 3-1. Environmental variable description in the Ohio database

3.2. Maryland's database:

TYPE OF DATA	NAME	DESCRIPTION	UNITS*
PHYSICAL, HABITAT AND MORPHOLOGIC PARAMETERS	Remoteness (REMOTE)	Rate based on the absence of human activity and difficulty of access	Score from 0 to 20*
	Shading (SHADING)	Rate based on estimates of the degree and duration of shading during the summer	Percentage *
	Epifaunal Substrate (EPI_SUB)	Amount of variety of hard, stable substrates usable by benthic macroinvertebrates	Score from 0 to 20*
	Instream habitat (INSTRHAB)	Perceived value of habitat to the fish community	Score from 0 to 20*
	Woody elements (WOOD)	Number of woody debris and rootwads in the control site	Number*
	Bank Stability (BANKSTAB)	Presence/absence of riparian vegetation and other stabilizing bank materials.	Score from 0 to 20*
	Velocity-depth diversity (VEL_DPTH)	Variety of velocity-depth regimes present at the site	Score from 0 to 20*
	Pool quality (POOLQUAL)	Variety and spatial complexity of slow or still water habitat	Score from 0 to 20*
	Riffle Quality (RIFFQUAL)	Depth, complexity and functional importance of riffle/run habitat	Score from 0 to 20*
	Channel alteration (CHAN_ALT)	Measure of large scale changes in the shape of the stream channel	Score from 0 to 20*
	Embeddedness (EMBEDDED)	Fraction of surface area of larger particles surrounded by fine sediment	Percentage*
	Channel Flow Status (CH_FLOW)	Fraction of the area of the stream that is covered by water	Percentage
	Aesthetics (AESTHET)	Visual appeal of the site and presence/absence of human refuse	Score from 0 to 20*
	Max. depth (MAXDEPTH)	Maximum depth at the site	Centimeters
	Riparian buffer width (RIP_WID)	Width of the riparian strip along the stream	Meters*
	Gradient (ST_GRAD)	Stream gradient	Percentage
	Average width (AVGWID)	Average wetted width	Meters
	Average thalweg (AVGTHAL)	Average thalweg depth	Centimeters
	Average velocity (AVG_VEL)	Average velocity	Meters per second
	Urban land use (URBAN)	Fraction of urban land use in drainage area	Percentage
	Forest, wetland, water land uses (FORWETWAT)	Fraction of unimpacted land uses in drainage area	Percentage
	Agricultural and barren land uses (AGRIBARR)	Fraction of agriculture/bare soil in drainage area	Percentage
	Drainage area (ACREAGE)	Catchment area at the site	Acres

TYPE OF DATA	NAME	DESCRIPTION	UNITS
CHEMICAL PARAMETERS	Temperature_FLD (TEMP-FLD)	Water temperature	Degrees Celsius
	Dissolved Oxygen_FLD (DO_FLD)	Dissolved oxygen	ppm
	pH_FLD (PH_FLD)	pH in summer time	Standard units
	Conductance_FLD (COND_FLD)	Specific conductance in summer time	µmho/cm
	Dissolved Organic carbon (DOC_LAB)	Dissolved organic carbon concentration	mg/L
	Nitrate (NO3_LAB)	Nitrate-Nitrogen concentration	mg/L
	Sulfate (SO4_LAB)	Sulfate concentration	mg/L

Table 3-2. Description of the environmental variables included in the MBSS database.

***The scoring system shown in the table corresponds to the old PHI. The scores for the new metrics were calculated with the guidelines from Paul et al. (2003) using the metrics in the old PHI**

3.3. Minnesota's database:

Type	Name	Description	Units
Habitat metrics and physical parameters	**Score Riparian**	QHEI metric	0 to 15
	MbufferWidth	Buffer width	Meters
	MBankEros	Bank Erosion	Percentage
	Score Substrate	QHEI metric	0 to 27
	PctEmbed	Embeddedness	Percentage
	Fines depth	Mean depth of fines	Cm
	PctRock	% of coarse substrates in transect	Percent
	PctBoulder	% of cover made of boulders	Percent
	Pctfine	% of fine substrate in transect	Percent
	PctPoolRun	% of reach that's pool and run	Percent
	PctRiffle	% of reach that is riffle	Percent
	PctRun	% of reach that is run	Percent
	PctPool	% of reach that is pool	Percent
	Score Cover	QHEI metric	0 to 17
	PctEmerMac	% of cover that is emergent macrophytes	Percent
	PctSubMac	% of cover that is submerged macrophytes	Percent
	PctWoody	% of cover that are woody elements	Percent
	PctOverVeg	% of cover that is overhanging vegetation	Percent
	PctOtherCov	% of cover that is other cover	Percent
	PctUnderCut	% of cover that is undercut	Percent
	PctCover	% cover for fish	Percent
	Score Channel	QHEI metric	0 to 36
	MWidth	Mean width	Meters
	MthalDepth	Maximum thalweg depth	Cm
	MDepth	Mean water depth at transect points	Cm
	Sinuosity	Ratio between stream length and straight distance	Ratio
	WDRatio	Width-depth ratio	Ratio
	Score Land use	QHEI metric	0 to 5
	PctDistLU	% disturbed land use in DA	Percentage
	PctUnDistLU	% undisturbed land use in DA	Percentage
	PctDistLU30	% disturbed land use in 30-meter buffer	Percentage
	PctUnDistLU30	% undisturbed land use in 30-meter buffer	Percentage
	DA	Drainage area	Sq. miles
	Gradient	Site slope	m/Km
Chemical parameters	Cond	Specific Conductance	
	DO	Dissolved oxygen	mg/L
	NH4	Ammonia	mg/L
	Nitrogen	Total nitrogen	mg/L
	pH	pH	Standard Units
	Phosphorus	Total phosphorus	mg/L
	Temp	Temperature	Degrees Celsius
	TSS	Total Suspended Solids	mg/L
	Turbid	Turbidity	

Table 3-3. Physical and chemical environmental variables used for clustering in the Minnesota database

4. Methodology

4.1. Obtaining clusters with similar characteristics

Because of the vast amount of data, we considered that clustering was necessary in order to reduce the natural environmental variability and identify sites that have similar in stream and off stream characteristics. SOM were used for this purpose. SOM were first developed by Kohonen (1991). They consist of a topologically ordered mapping of the input space (in our case multiple environmental variables) onto a two-dimensional grid with a meaningful order. All the input parameters are located following a weighting algorithm onto different grid sites (called neurons or cells) according to their similarity with the neighboring cells. Therefore, similar groups of data or clusters are easily identified. SOM have been widely used in different fields such as speech recognition or economics, and are now being discovered as a great tool for environmental purposes (Brosse et al., 2001, Virani et al., 2005).

In our case, we used the SOM applet *"somtoolbox"* in MATLAB 7.1 in order to perform the clustering. The clustering was performed using all the records shown in the database description. Since the data correspond to either in stream and off stream physical environmental variables or water quality measurements, the SOM should identify clusters with similar physical and chemical characteristics, which means that those sites have similar stressors to the biotic community. Once the SOM were run, the distribution within clusters of the biotic indices (fish, macroinvertebrate or both if available) were plotted to check if sites with similar type of stressors (clusters) were translated into similar biotic integrity in those sites. The habitat indices for each state were also plotted to identify the importance of habitat over the biotic community and separate habitat/physical stressors effects from chemical impairment effects.

4.2. Environmental variable selection for prediction purposes

The metric selection was performed in two different ways. The first way was using the metrics that were mainly responsible for the differences among the SOM clusters. MRT were used for this purpose. The metrics whose cluster distribution followed a similar pattern to the biotic indices distribution were considered to be the most important. The second way for variable selection was using PCCA for the whole state dataset in order to identify those variables that have an overall deeper impact on the fauna. The same number of metrics was used with both methods to compare their performances for prediction purposes.

Multiple Range Tests (MRT)

MRT consist of comparisons between different groups of data. The test identifies homogeneous groups and analyzes the differences among each group's mean using Fisher's Least Significance Difference (LSD). Fisher's LSD then determines if the differences within groups are statistically significant. This test was run using the Software *Statgraphics 5 Plus.*

The MRT were run in each state for the biotic indices (fish or macroinvertebrate) to test if the biotic integrity within the clusters found with the SOM were statistically significant. Subsequently, the MRT were run for each one of the environmental variables used in the SOM. The tests were based on the assumption that those parameters with a strong effect on the biotic community will most likely have the same (positive effect), opposite (negative effect) or similar cluster significant differences among homogeneous groups than the biotic indices. The metrics that showed similar homogeneous groups' distributions were selected for biotic integrity prediction. The number of variables selected was not set and depended on how many parameters showed similar distributions in each case.

Polynomial Canonical Correspondence Analysis (PCCA)

Canonical Correspondence Analysis (CCA) is a widely used exploratory technique able to identify different roles of the explanatory variables (environmental variables) over the response variables (fish or fish species counts). CCA was initially developed by Ter Braak (1986). CCA is an extension of correspondence analysis. However, CCA imposes the extra restriction that the canonical axes have to be linear combinations of the environmental variables. CCA has been successfully used in many cases in order to identify the effects of environmental variables over the biotic community (Fernandez et al., 1998, Bhat, 2004, Belore et al., 2002).

One of the main problems that traditional CCA presents is that, even though a chi-square transformation of the transformed variables is done, the relationship between the transformed response data and the explanatory variables is still assumed to be linear. PCCA is based on the same principles as CCA but with the key difference that the regressions are performed with highly non-linear equations. There's no reason why nature should linearly relate changes in species assemblages to changes in environmental variables (Makarenkov and Legendre, 2002).

In order to run the PCCA, two different matrices were constructed for each state. The explanatory variables (X) which included all the variables shown in Tables 4.1, 4.2 and 4.3 and the response variables (Y), which included the individual fish or fish species counts corresponding to each one of the metrics included in the states' fish IBI, when available (not available in Minnesota). Matrix X values were logged (natural Log) and centered in their means before running the PCCA. A freeware available on the Internet developed by Makerenkov and Legendre (2002) was used to perform the analysis. Once the model was run, the site scores were represented on a two-axis plot along with the environmental gradients with a MATLAB 7.1 application developed by our team.

In order to identify the variables with biggest effect over biotic community, the distance between the origin and the projected site points over each one of the environmental gradients was measured. This operation was performed for each point considering positive the points that fell on the same side as the environmental gradient and negative otherwise. Subsequently, all the points' projections over each environmental variable were obtained, averaged and then ranked based on the absolute value of the average distance between the points and the origin. The variables with largest absolute values were considered to be the ones with the deepest impact on the biotic community. The projection methodology is shown in Figure 4-1. This procedure is based on the CCA interpretation guidelines given by Jongman et al. (1995). The PCCA analysis was run for the entire dataset in each state, not on a cluster basis.

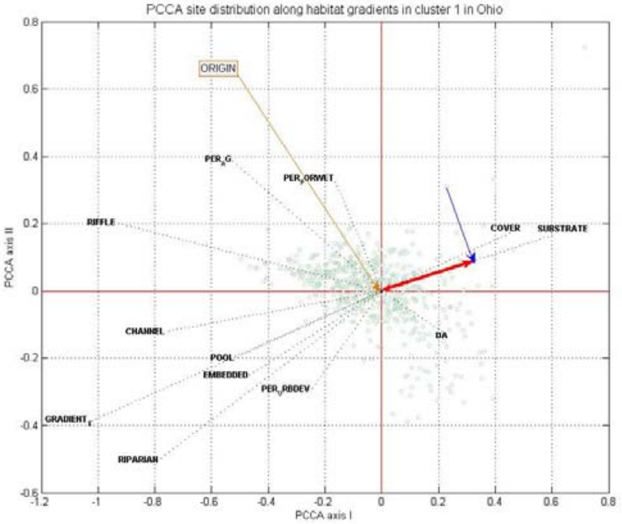


Figure 4-1. Interpretation of the plots obtained with the PCCA for the ranking of environmental variables. In this case one point is being projected over the substrate score and the distance from the origin is measured

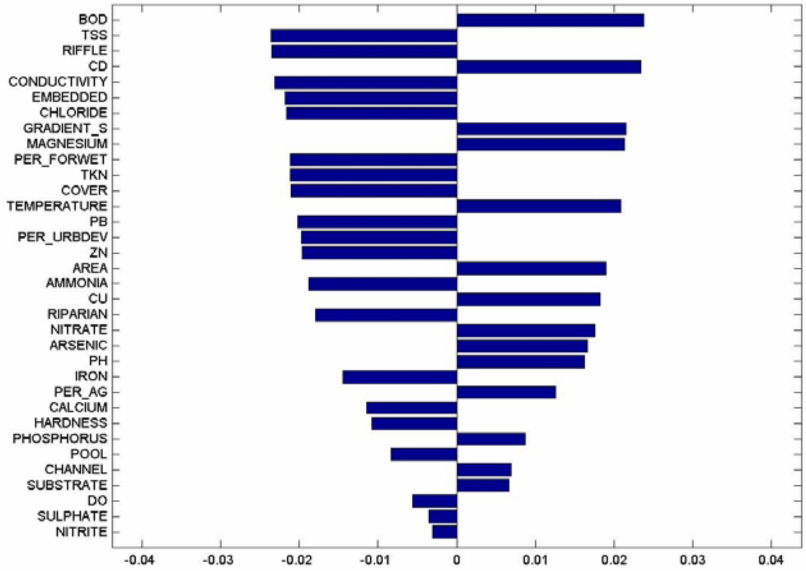


Figure 4-2. Example of environmental variable ranking. Right side means positive correlation between the variable and the entire biotic community, left side is means the opposite

4.3. Regression methodology

The regression methodology used was the same method that Makarenkov and Legendre (2002) used for the PCCA. The metrics selected in the SOM+MRT or PCCA were regressed versus the fish IBI (Benthic IBI in Maryland). Permutation tests of significance for the regression equations were performed and considered significant if $p<0.05$. The regression methodology consisted of four different steps:

1. Simple least-squares multiple linear regression of the dependent variate (y) on all variables in X

$y_c = Xb = X[X'X]^{-1}y$ **(Eq. 1)**.

where y_c are the calculated biotic indices, b are the regression coefficients, X is the matrix of independent variates and y is the matrix with biotic indices.

2. Residual calculation

$y_{res} = y - y_c$ **(Eq. 2).**

where y_{res} is the residual obtained in each site between observed and calculated index.

3. Residual calculation using the two independent variates that provide the best quadratic approximation

$y_{res}^{calc} = X^{ij}c^{ij}$ **(Eq. 3).**

where c^{ij} is a matrix with the vector regression coefficients for the explanatory variables. Matrix X^{ij} is constructed as follows:

$$X^{ij} = \begin{pmatrix} x_{1i} & x_{1j} & x_{1i}x_{1j} & x_{1i}^2 & x_{1j}^2 & 1 \\ x_{2i} & x_{2j} & x_{2i}x_{2j} & x_{2i}^2 & x_{2j}^2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{ni} & x_{nj} & x_{ni}x_{nj} & x_{ni}^2 & x_{nj}^2 & 1 \end{pmatrix}$$

4. A new combined environmental (t) variable is obtained using the following equation

$x_{mt} = x_{mi}b_i + x_{mj}b_j + y_{res,m}^{calc(i,j)}$ **(Eq. 4).**

where the coefficients b are those of. Eq.2. A new combined vector (X_{mt}) is created and the initial explanatory variables' vectors (X_i and X_j) are removed. These four steps are repeated until only one explanatory variable remains, in which every single variable of X is expressed by linear and quadratic terms in the final combined explanatory variable X($n \times 1$). The final values for the combined variable are obtained by a simple linear regression of y on X ($n \times 1$). The regression equation obtained was used to calculate the final calculated indices based on the explanatory variables regression model. In all the models run, the number of observations (n) has to be greater than the maximum number of independent and combined variables($3m$-1), where m is the number of independent variates, as advised by Makarenkov and Legendre (2002).For our regressions we imposed the extra condition that *(3m-1)* had to be equal or smaller than (*n/2).*

4.4. Comparison of observed and calculated values using a binning system

Biotic integrity is usually measured with an IBI index. Different IBI values mean different conditions or different instream biotic community health. However, often times, watershed managers are more interested on a range than the IBI value itself. For example, Maryland's MBSS (1999) considers four different levels for the biotic community integrity: good (IBI between 4 and 5), fair (between 3 and 3.9), poor (between 2 and 2.9), and very poor (between 1 and 1.9). The observed IBIs as well as the results from the regression equations were subdivided in four equal sized bins and compared. The number of times in which the predictive model was forecasting the right or wrong bins was recorded.

5. Results

5.1. Ohio

Clustering with the SOM

Three very prominent clusters were obtained with the SOM using all the variables shown in Table 3-1. The SOM grid as well as the fish IBI distribution among clusters is shown in the following figure (top line means 75th percentile, red line is median value and bottom line means 25th percentile).

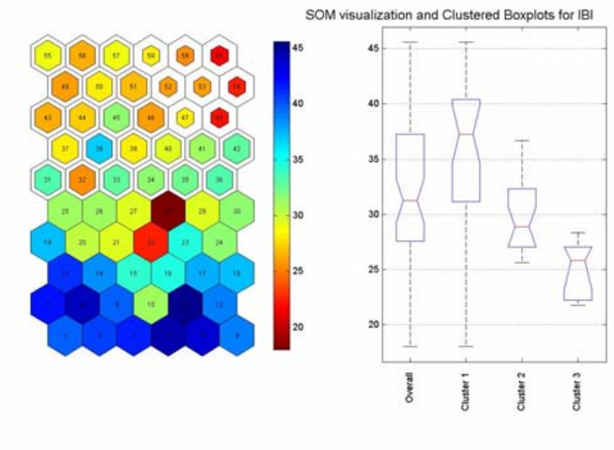


Figure 5-1. Fish IBI cluster distribution in Ohio

Environmental variables selection

MRT:

When the MRT test were run, the following parameters had the same distribution of homogeneous groups within clusters: PER_AG, AS, BOD, CHANNEL, COVER, EMBEDDEDNESS, PER_FORWET, GRADIENT_S, POOL, RIFFLE, RIPARIAN and SUBSTRATE. We also chose iron, ammonia and conductivity because these were significantly different in clusters 1 and 3.The complete MRT are included in the Appendices.

PCCA:

The ranking of the different environmental variables obtained with the PCCA analysis is shown in the following figure

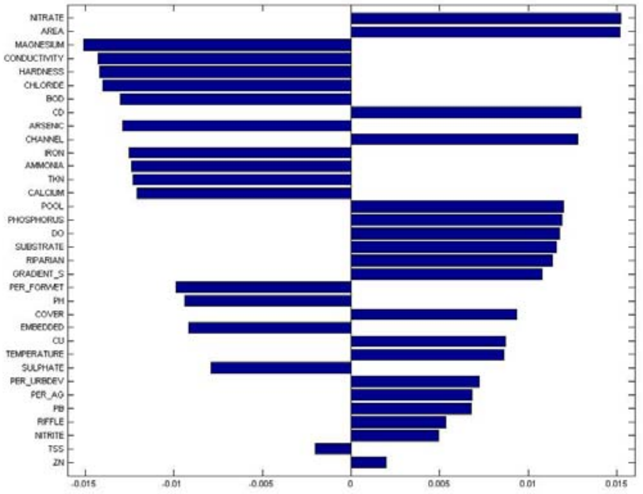


Figure 5-2. Environmental variable ranking in Ohio based on PCCA

The first two canonical axes in which the variable-ranking was based explained a 23% of the total variance in the Ohio dataset.

Biotic integrity prediction

MRT metrics

The results of the regression equations with the metrics selected with this methodology are shown in the following figure.

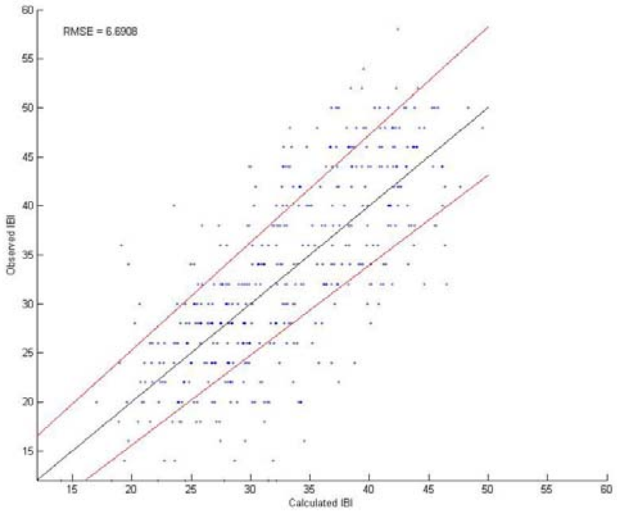


Figure 5-3. Fish IBI prediction in Ohio based on metrics selected from the MRT analysis

PCCA metrics

The top fifteen metrics were selected and the same procedure was applied. Results are shown in the following figure

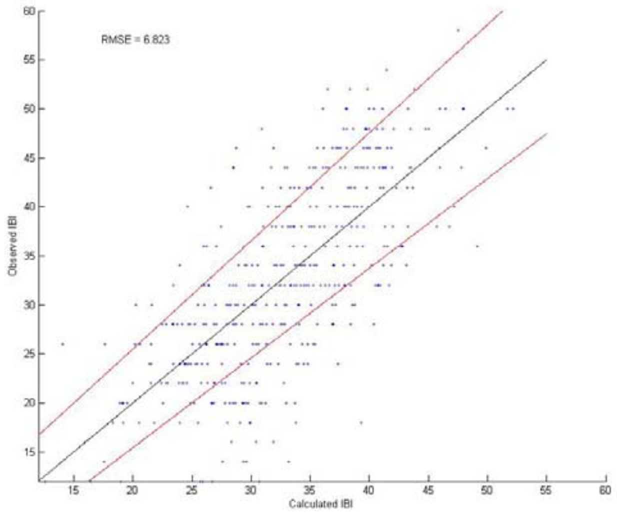


Figure 5-4. Fish IBI prediction in Ohio based on metrics selected from the PCCA analysis

Data binning

The bin predictions are shown in the following figures

MRT metrics

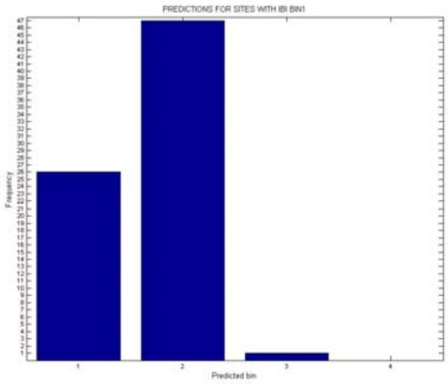


Figure 5-5. Predictions for bin 1

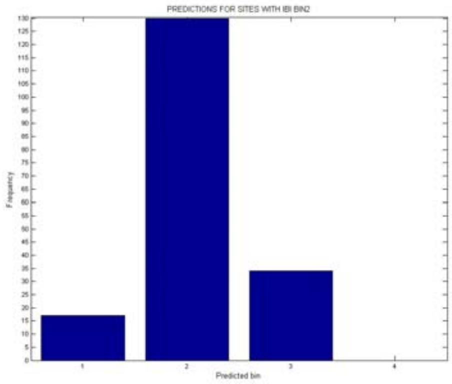


Figure 5-6. Predictions for bin 2

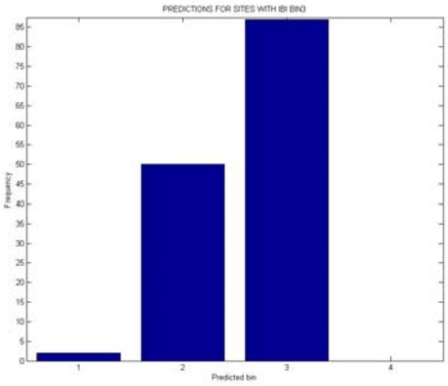


Figure 5-7. Predictions for bin 3

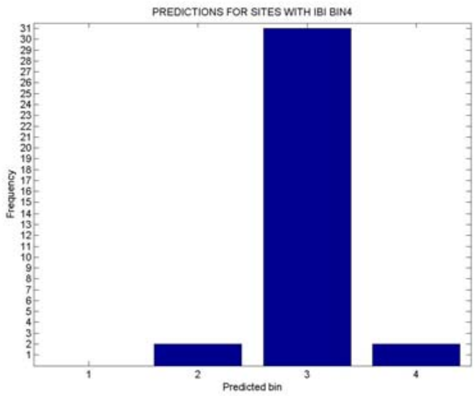


Figure 5-8. Predictions for bin 4

PCCA metrics

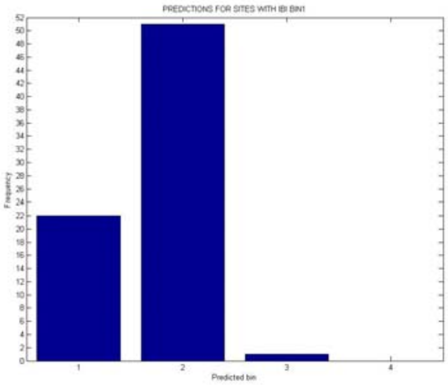


Figure 5-9. Prediction for bin 1

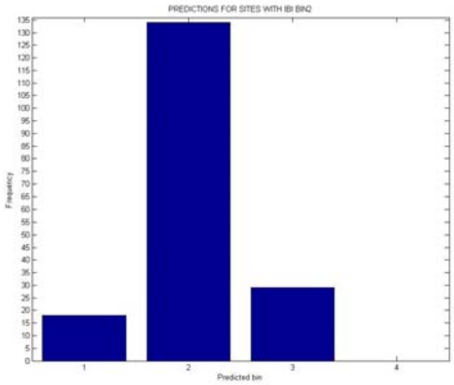


Figure 5-10. Predictions for bin 2

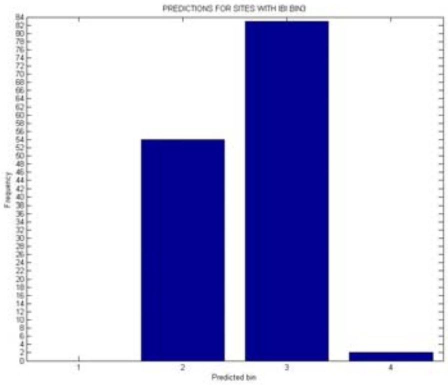


Figure 5-11. Predictions for bin 3

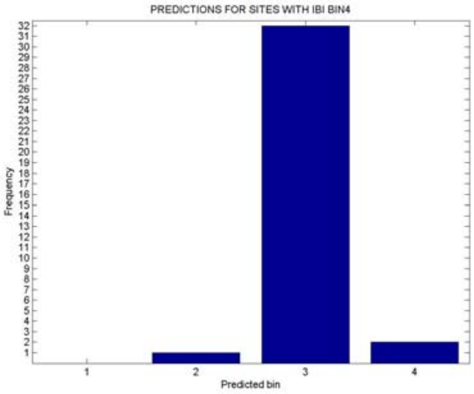


Figure 5-12. Predictions for bin 4

The following table is a summary of the predictions in Ohio state

METHOD	# OF SITES	r	RMSE	p	% times in right bin
MRT	428	0.73	6.69	0.0099	57
PCCA	428	0.71	6.82	0.0099	63

Table 5-1. Summary of the model performance in Ohio

5.2. Maryland

5.2.1. Coastal sites

Clustering with the SOM

Five clusters were detected by the SOM when all the metrics were used. The SOM grids along with the benthic IBI cluster distributions are presented in the following figure. In this case we used the benthic IBI instead of the fish IBI because it is reported that fish IBI is biased with stream size (Southerland et al., 2005)

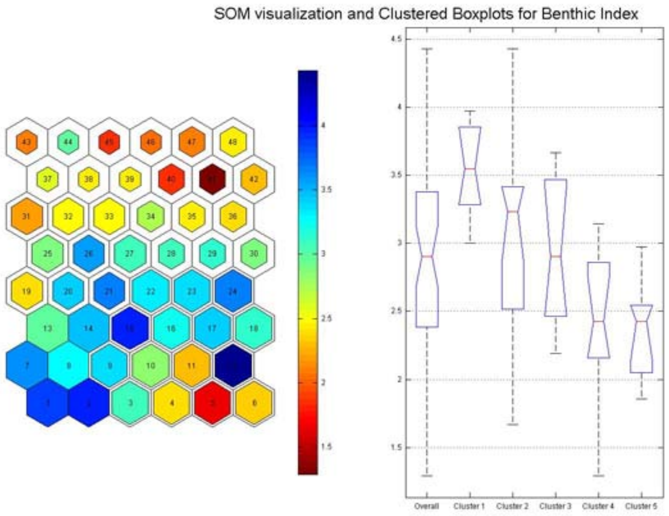


Figure 5-13. Benthic IBI cluster distribution in coastal sites in Maryland

Environmental variables selection

MRT:

The environmental variables that followed similar distribution to the benthic biotic index were: riffle quality, velocity-depth diversity, average velocity, epifaunal substrate, embeddedness, pool quality, riparian width, DOC, nitrate, water temperature, maximum depth, channel flow, and sulfate.

PCCA:

The ranked environmental variables from the PCCA analysis in Maryland's coastal areas clusters are shown in the following figures

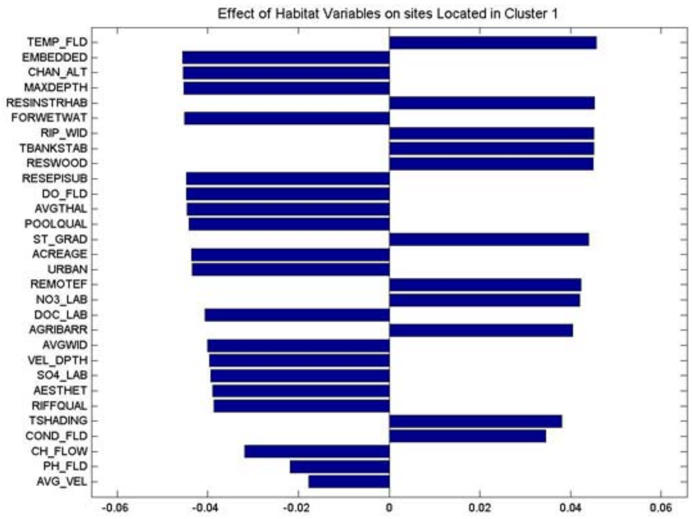


Figure 5-14. Environmental variable ranking in coastal sites in Maryland

The PCCA analysis in Maryland was run with the assumption that both, fish and benthic communities have similar responses to similar stressors. The PCCA is based on individual fish and fish species counts. However, the predictions in Maryland are performed for the benthic IBI for the reason stated previously.

Biotic integrity prediction

MRT metrics

Benthic IBI predictions were performed with the thirteen MRT selected metrics

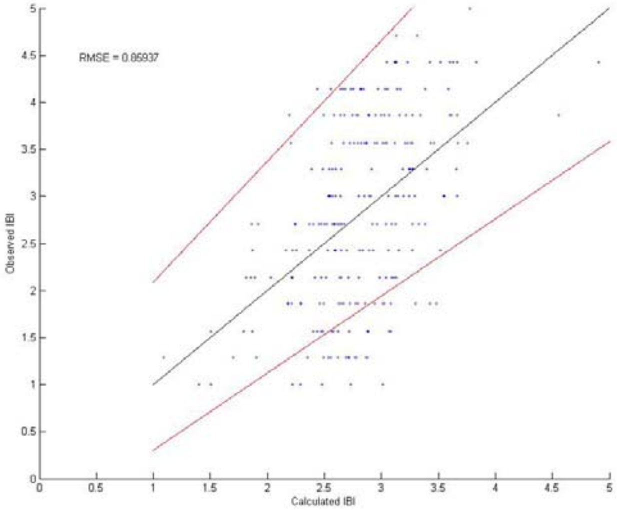


Figure 5-15. BIBI predictions using metrics from MRT

PCCA metrics

The top thirteen PCCA metrics were used for prediction

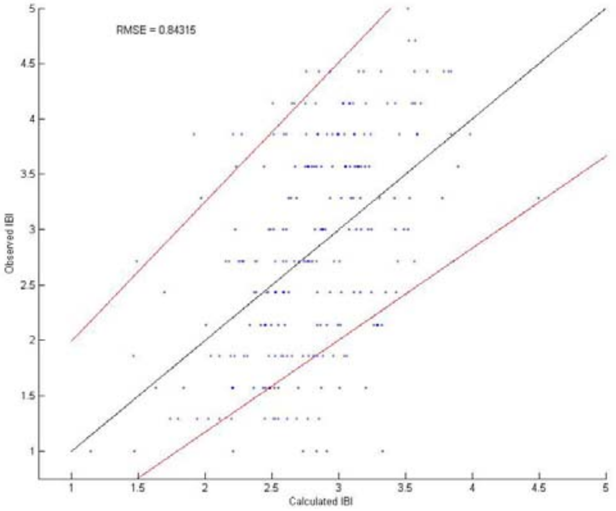
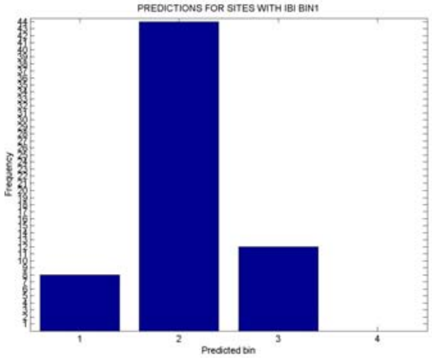


Figure 5-16. BIBI prediction using metrics from PCCA

Data binning procedure

MRT metrics



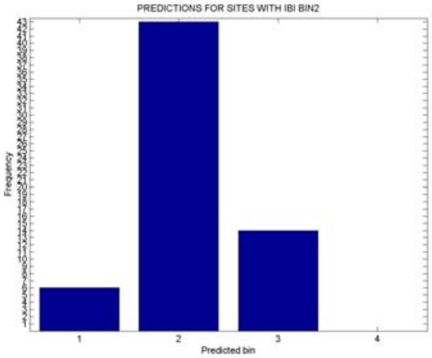


Figure 5-17. prediction for bin 1

Figure 5-18. Predictions for bin 2

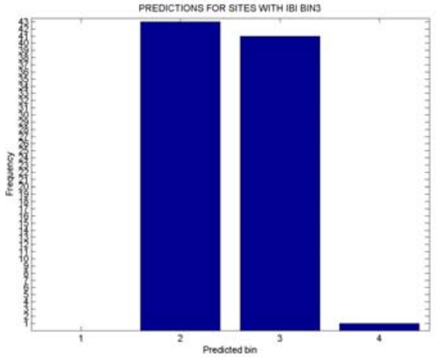


Figure 5-19. Predictions for bin 3

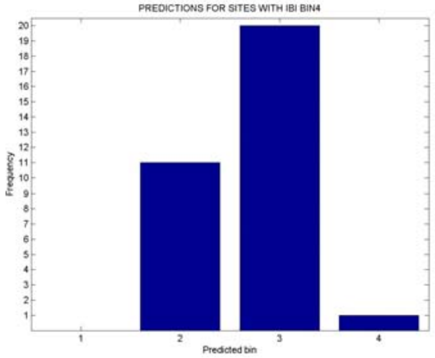


Figure 5-20. Predictions for bin 4

PCCA metrics:

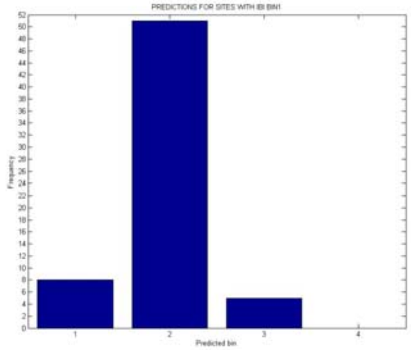


Figure 5-21. Predictions for bin1

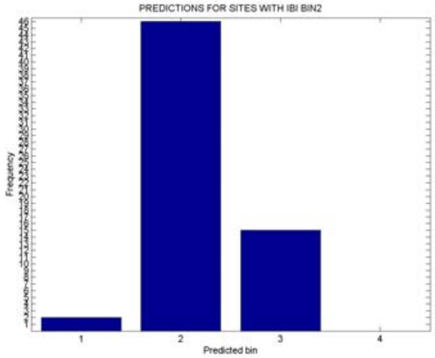


Figure 5-22. Predictions for bin 2

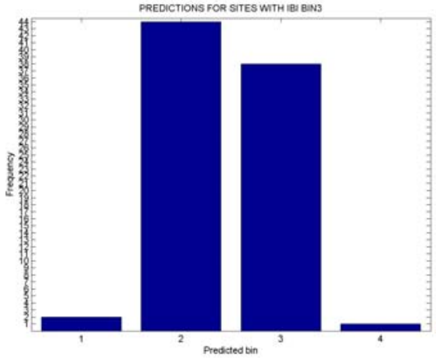


Figure 5-23. Prediction for bin 3

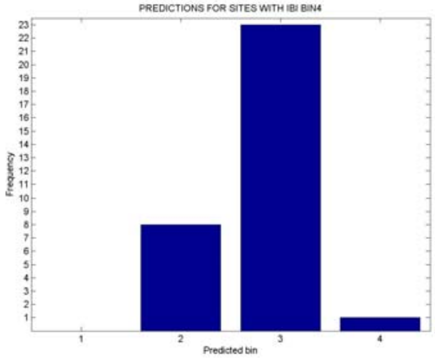


Figure 5-24. Predictions for bin 4

This table summarizes the results in Maryland coastal sites:

METHOD	# OF SITES	r	RMSE	p	% times in right bin
MRT	244	0.50	0.86	0.14*	38
PCCA	244	0.52	0.84	0.04	38

***Not statistically significant (p>0.05)**

Table 5-2. Regression parameters for the fish IBI predictions in coastal sites in Maryland

5.2.2. Piedmont sites

Clustering with the SOM

The benthic IBI cluster distribution in Piedmont areas is shown in the following figure

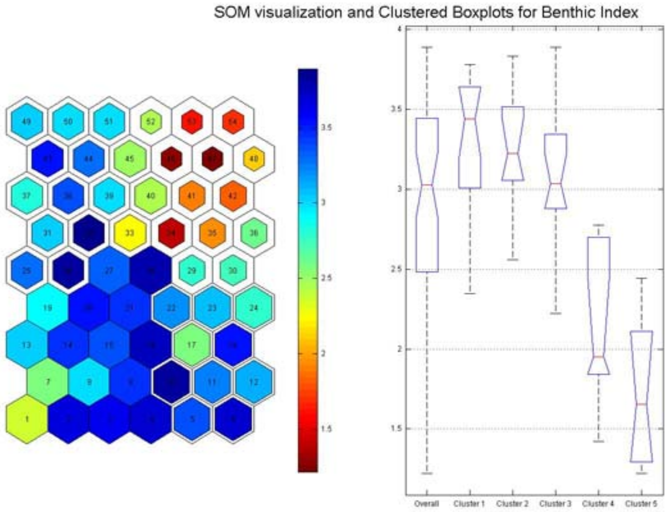


Figure 5-25. Benthic IBI cluster distribution in Piedmont sites in Maryland

Environmental variables selection

MRT:

The environmental variables with similar distributions within clusters were: agricultural land use, urban land use, aesthetic quality, remoteness, specific conductance, pH, nitrate and sulfate concentrations. DOC was also included as a discriminating metric.

PCCA:

The metrics ranking from the PCCA analysis is shown in the following plot

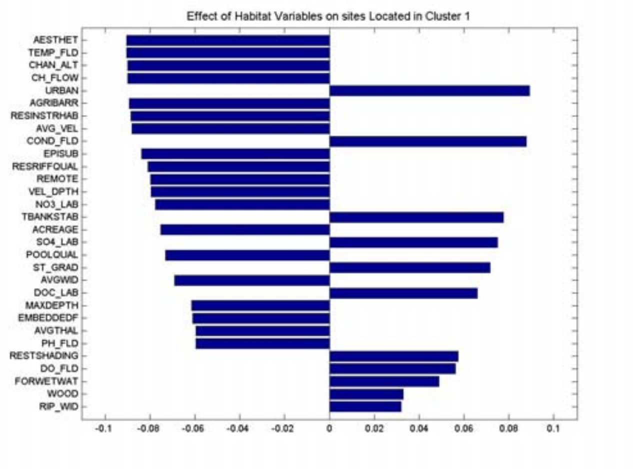


Figure 5-26. Environmental variable ranking in Piedmont sites in Maryland

Biotic integrity prediction

MRT metrics

The regression with the nine selected metrics is shown as follows

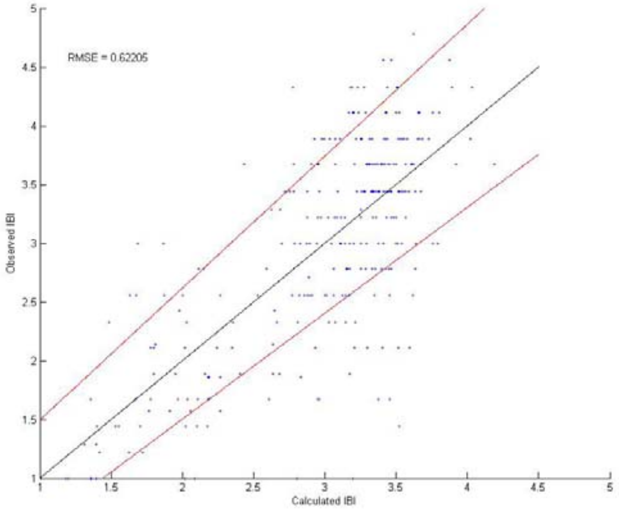


Figure 5-27. BIBI prediction using metrics from the MRT analysis

PCCA metrics

The regressions with the top nine metrics is as follows

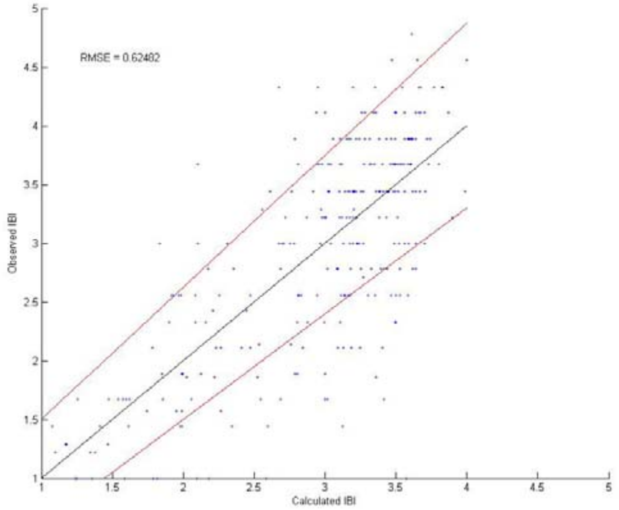


Figure 5-28. BIBI prediction using metrics from PCCA

Data binning

MRT metrics

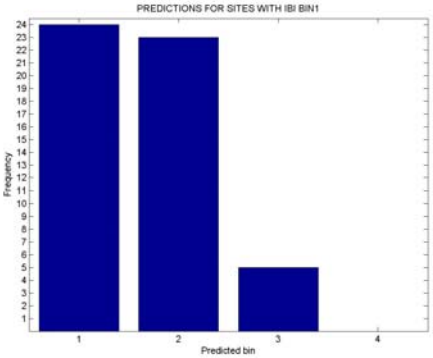


Figure 5-29. Prediction for bin 1

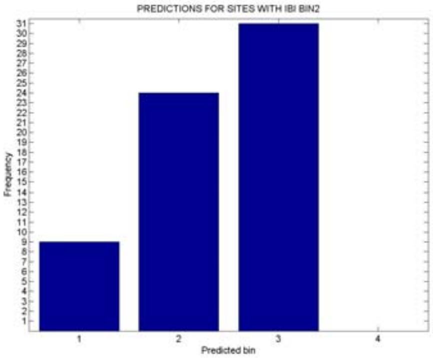


Figure 5-30. Prediction for bin 2

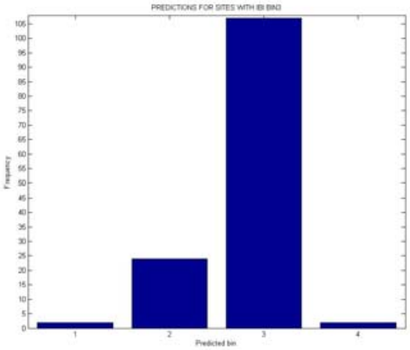


Figure 5-31. Prediction for bin 3

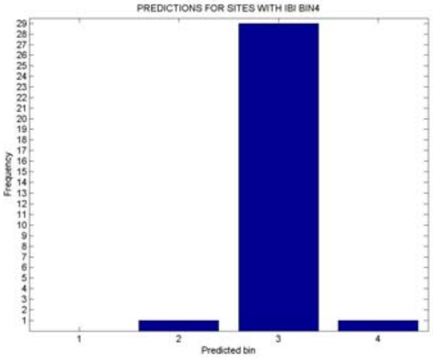


Figure 5-32. Prediction for bin 4

PCCA metrics

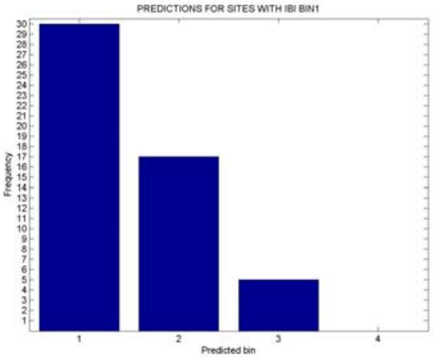


Figure 5-33. Prediction for bin 1

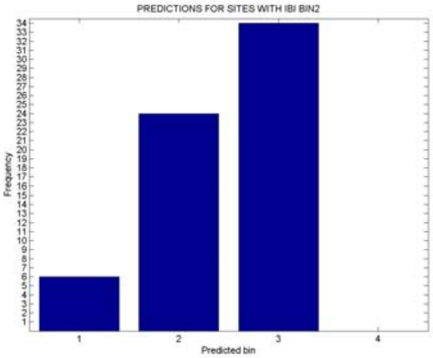


Figure 5-34. Prediction for bin 2

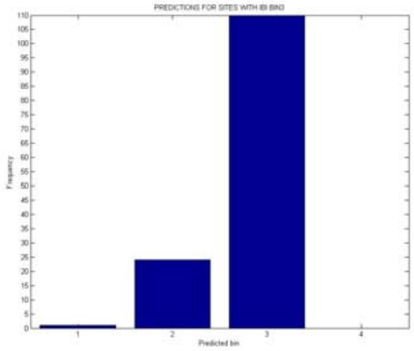


Figure 5-35. Prediction for bin 3

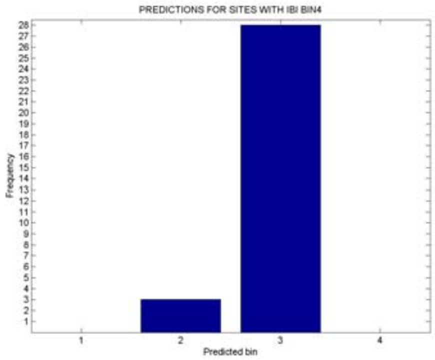


Figure 5-36. Prediction for bin 4

The following table summarizes the results in Maryland's Piedmont sites

METHOD	# OF SITES	r	RMSE	p	% times in right bin
MRT	282	0.73	0.62	0.0099	55
PCCA	282	0.73	0.62	0.0099	58

Table 5-3. Regression parameters for the fish IBI predictions in piedmont sites in Maryland

5.2.3. Highland sites

Clustering with the SOM

The benthic IBI distribution among clusters is shown in the following figure:

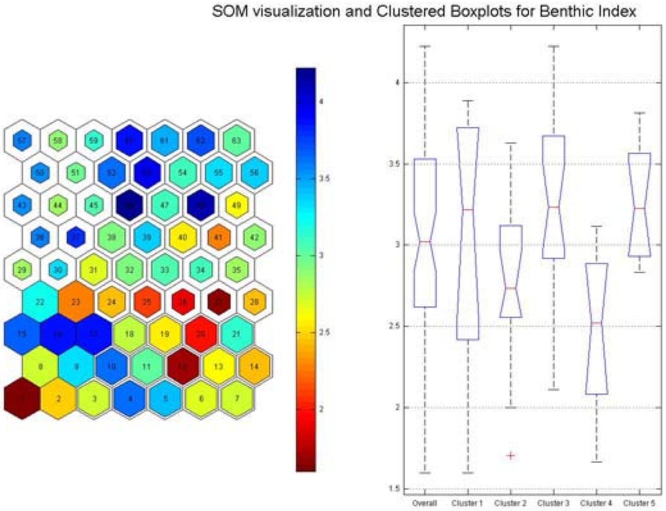


Figure 5-37. Benthic IBI cluster distribution in Highland sites in Maryland

Predicting environmental variables selection

MRT

The metrics that best matched the distribution of the BIBI among clusters were: agriculture and forest land uses, bank stability, remoteness, riparian width, aesthetic quality, conductance, nitrate, pH and temperature.

PCCA

The variables ranking based on the PCCA analysis is shown as follows:

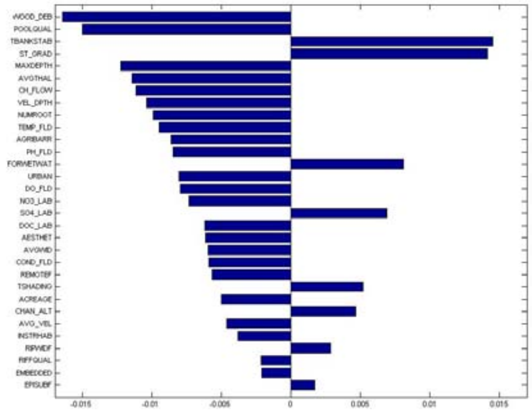


Figure 5-38. Environmental variable ranking based on the PCCA analysis

Biotic integrity prediction

MRT metrics

The regression obtained is as follows

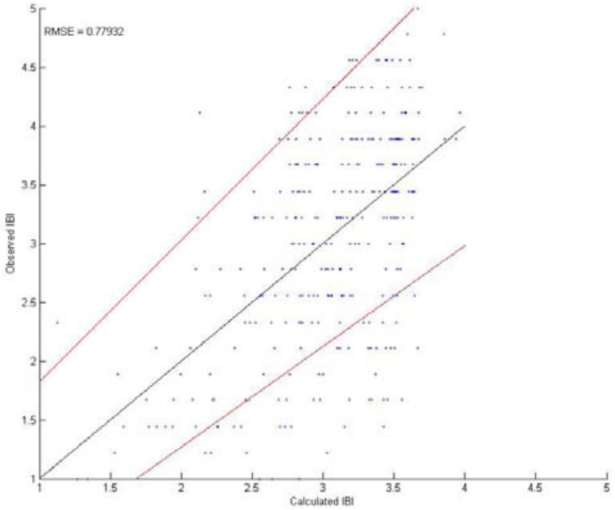


Figure 5-39. BIBI prediction using metrics from the MRT analysis

PCCA metrics

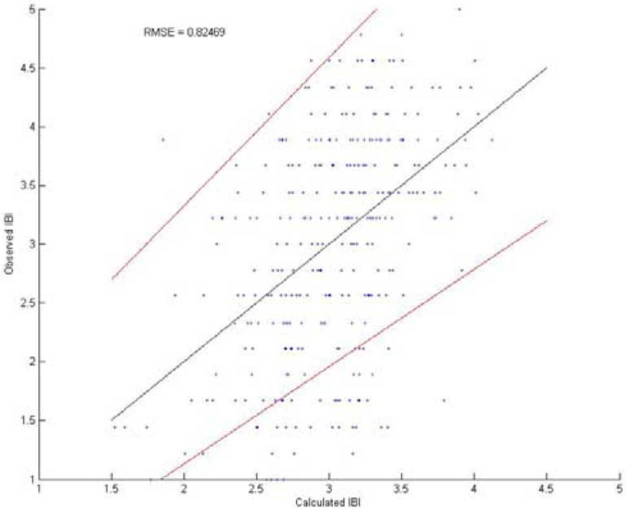


Figure 5-40. BIBI prediction using metrics from PCCA

Data binning

MRT metrics

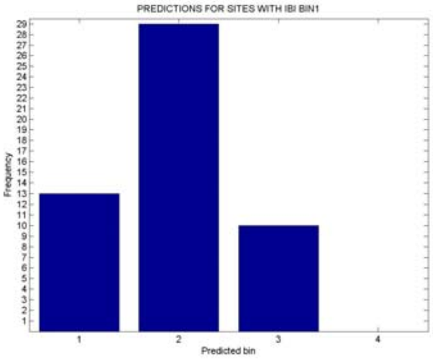


Figure 5-41. Prediction for bin 1

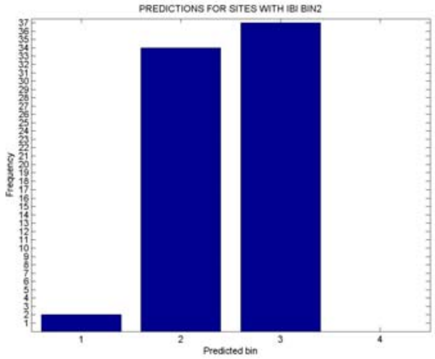


Figure 5-42. Prediction for bin 2

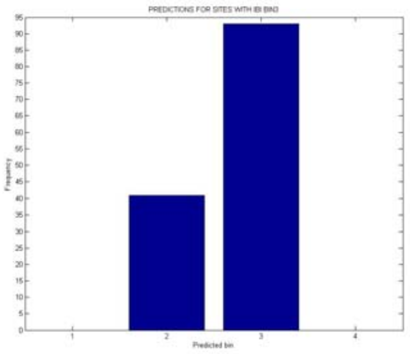


Figure 5-43. Prediction for bin 3

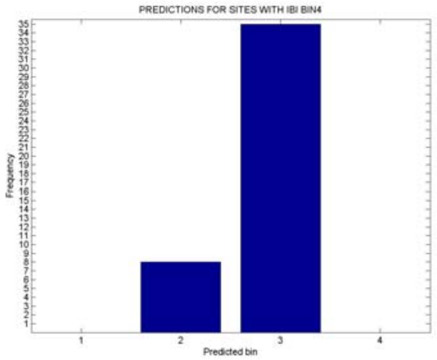


Figure 5-44. Prediction for bin 4

PCCA metrics

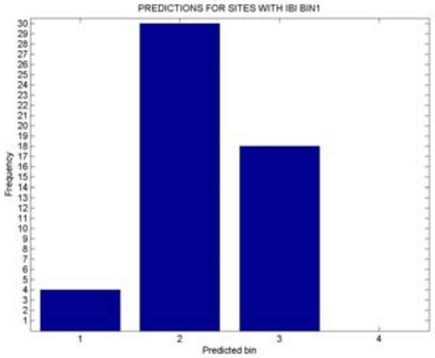


Figure 5-45. Prediction for bin 1

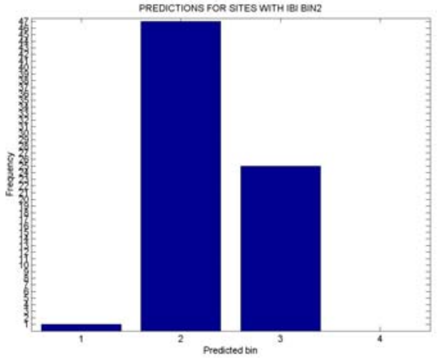


Figure 5-46. Prediction for bin 2

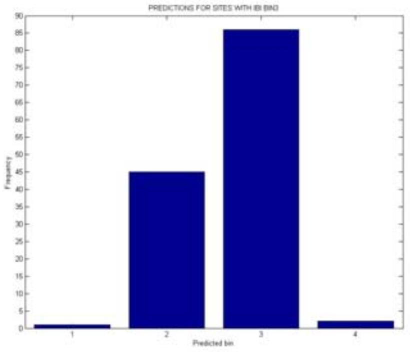


Figure 5-47. Prediction for bin 3

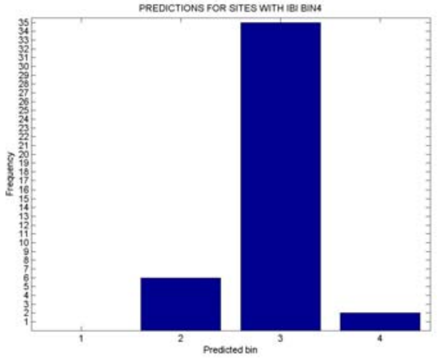


Figure 5-48. Prediction for bin 4

The next table summarizes the results in Maryland's highland sites

METHOD	# OF SITES	r	RMSE	p	% times in right bin
MRT	302	0.55	0.78	0.0099	46
PCCA	302	0.48	0.82	0.0099	46

Table 5-4. . Regression parameters for the fish IBI predictions in highland sites in Maryland

5.3. Minnesota

Clustering with the SOM

Minnesota's database was clustered using all the available environmental variables. The fish IBI cluster distribution and the SOM grid are shown as follows.

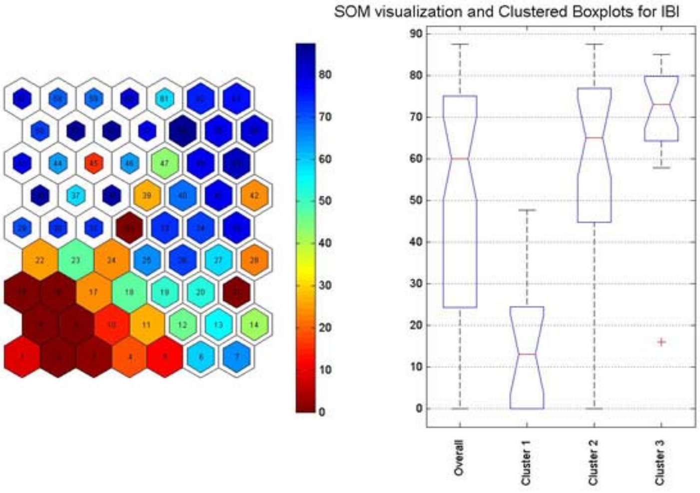


Figure 5-49. Fish IBI distribution among clusters in Minnesota

Predicting environmental variables selection

As shown in Table 3-3, Minnesota's database had records for the five QHEI metrics scores as well as the actual measurements that comprise these scores (i.e. percent of boulders in substrate). Two different types of predictions were performed. The first one included the QHEI scores, DA, gradient and chemical parameters as predicting metrics, the second one was the same but the actual measurements were taken as explanatory variables. The PCCA analysis could not be performed because the fish counts were not available.

Biotic integrity predictions with QHEI scores

MRT

The variables selected from the MRT were: land use score, riparian score, cover score, channel score, conductance, nitrogen, pH, phosphorus and TSS.

Biotic integrity predictions with actual measurements

MRT (most discriminant measurements)

Due to the high number of parameters, when several variables had high correlation (r>0.8), only one was selected and the rest excluded. DA was also excluded. The variables used were: percent disturbed land use in 30-meter buffer, bank erosion, percent embeddedness, percent rock, percent boulder, percent run, percent riffle, percent pool, percent over vegetation, percent woody elements, width-depth ratio, mean depth, gradient, conductance, pH, and TSS.

MRT (only substrate, morphologic and land use parameters)

In this case the measurements used were: percent disturbed land use in 30-meter buffer, percent embeddedness, percent rock, percent boulder, percent pool-run, percent riffle, width-depth ratio, mean depth, and gradient.

MRT (only substrate, morphologic and land use parameters plus two water quality patameters)

In this case the measurements used were: percent disturbed land use in 30-meter buffer, percent embeddedness, percent rock, percent boulder, percent pool-run, percent riffle, width-depth ratio, mean depth, and gradient, pH and conductivity.

Biotic integrity preidction

Using QHEI scores

The regression is shown as follows

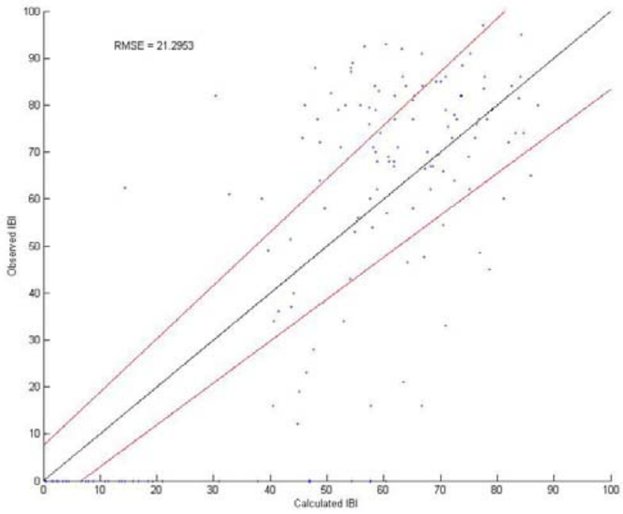


Figure 5-50. Fish IBI prediction with QHEI scores

Using actual measurements

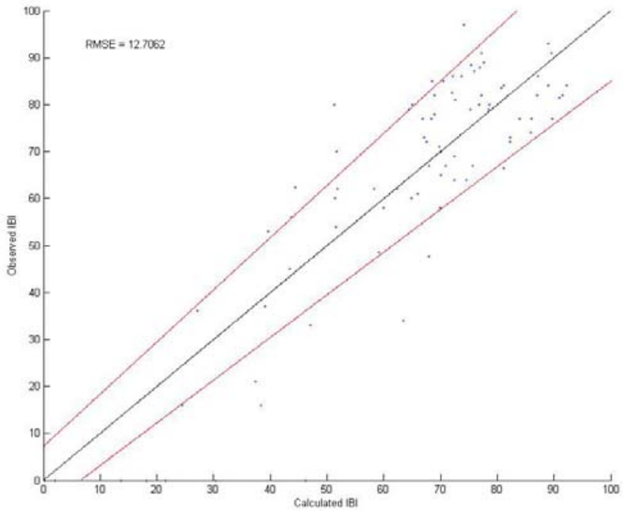


Figure 5-51. Fish IBI prediction with actual measurements

Using only substrate, morphology and land use measurements

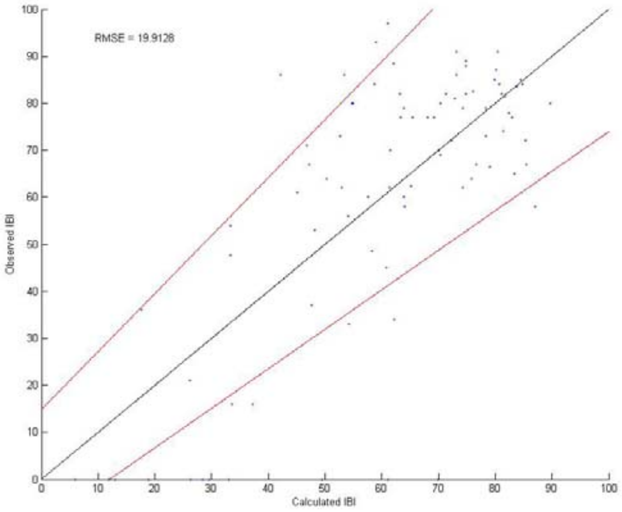


Figure 5-52. Fish IBI prediction with substrate, morphologic and land use parameters

Using substrate, morphology, land use and pH and conductivity

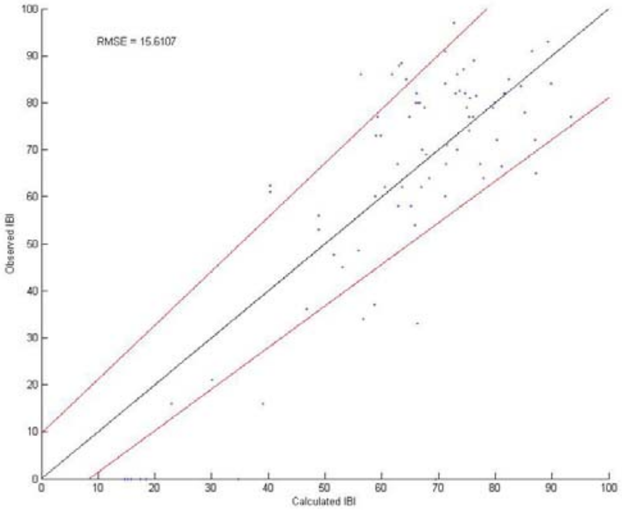


Figure 5-53. Fish IBI prediction with substrate, morphologic, land use and two water quality parameters

Data binning

The results for the binning procedure as well as the summary for the regressions in each case are shown in the following table.

METHOD	# OF SITES	r	RMSE	p	% times in right bin
QHEI scores	162	0.79	21.3	0.0099	53
Actual measurements	88	0.91	12.7	0.0099	67
Subs+morph+LU	88	0.75	19.91	0.02	41
Subs+morph+LU+WQ	88	0.82	15.61	0.0099	53

Table 5-5. Regression parameters for the fish IBI predictions in Minnesota

6. Conclusions and future work

The regression methodology presented in this report proved to be promising, and worked fairly well especially in the states of Ohio and Minnesota. These two states' inter-cluster biotic integrity could be mainly determined by habitat quality rather than water quality as determined by the metric selection with the MRT tests in which practically all the habitat metrics were chosen as discriminant parameters. However, this doesn't mean that water quality is not important as shown in the PCCA test in Ohio and the MRT tests in Minnesota.

In the case of Maryland, the model only seems to work well in piedmont sites. In coastal and highland sites it worked poorly. Some reasons could explain the difference in performance. In first place, the SOM were run using all the habitat and water quality values that were available. In each of the three strata (coastal, piedmont and highland) the habitat index is calculated differently using different metrics (see Paul et al., 2003). This means that when the clustering was performed, some physical habitat metrics were part of the new PHI, while some others corresponded to the old PHI. The metrics from the old PHI were based on reference sites that were found looking at their biotic integrity. Since Maryland's old fish IBI (the one we had in our database) is known to be biased with stream size, the old habitat metrics are also biased (Southerland et al., 2005). In second place, the predictions were done for benthic IBI instead of fish IBI because the benthic IBI doesn't have this problem. Since the old habitat metrics are based on biotic integrity based on fish IBI, the correlation between habitat characteristics and benthic community is not necessarily clear. Also, in Maryland, the new PHI metrics might not be very discriminant since only one was selected for the coastal sites' predictions with the MRT methodology, and only one in piedmont sites (whose biotic integrity seems to be more linked to water quality and land use patterns than habitat according to the MRT).

Both metric selection metrics (PCCA and SOM+MRT) seem to work very similarly. The selection of metrics using MRT usually yields higher *r* values. However, the bin predictions are usually better with the PCCA. This could mean that even though PCCA's regressions are poorer overall, it is able to better identify the variables that mainly determine biotic integrity distribution. One drawback in the case of the PCCA is that it only accounts for the first two canonical axes, which accounts for only a part of the total variability.

The use of actual measurements instead of scores seemed to work very well in the case of Minnesota. The most important parameters identification with the SOM+MRT technique and elimination of highly correlated data (r>0.8) proved to be a good tool for data selection. In the case of Minnesota, the differences between cluster 2 and 3 were mainly due to substrate and morphologic habitat parameters. The differences between cluster 1 and the rest were due to habitat and water quality. The selection of actual substrate and morphologic measurements along with only two discriminant water quality parameters resulted in a good prediction of biotic integrity. This proves that for prediction purposes, actual measurements instead of habitat scores might work better. By identifying those features that truly affect biotic community, we can accurately predict it.

The data binning method yielded 50% of correct predictions in average. One problem arises with this methodology. Usually some part of the data are underrepresented (i..e usually there's less data points in extremes or bins 1 and 4). Since the main goal of the regression methodology is obtaining the highest possible r^2, sites located in a bin with a lot of data points will be better predicted than others. Predictions for bin2 and 3 are usually better than those for bins 1 and 4 because is more common to find sites with fair/poor (bin2) or fair/good (bin3) biotic integrity than sites with poor/very/poor (bin1) or good/excellent (bin4) biotic integrity.

As future research in order to fine-tune the model, a data selection for each state should be performed. The datasets should contain equal number of sites in each bin in order to avoid the problems just described. Also, removal of outliers is important. Even though it is possible (albeit unlikely) to find poor biotic integrity with excellent habitat and water quality, it is not common. Sites with unusual characteristics should be removed because the main goal for the model is to predict correctly as many times as possible. The presence of outliers may substantially change the regression equations and jeopardize the overall model performance. In the case of Maryland, the SOM should be run again using only the new habitat metrics and water quality parameters to avoid the problems cited previously.

7. References

Barbour, M.T., J.Gerritsen, B.D. Snyder and J.B. Stribling (1999). Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington D.C.

Belore, M.L., Winter, J.G., Duthie, H.C. (2002). Use of diatoms and macroinvertebrates as bioindicators of water quality in Southern Ontario Rivers. Canadian Water Resources Journal, Vol. 27(4):457-484

Bhat, A.(2004). Patterns in the distribution of freshwater fishes in rivers of Central Western Ghats, India and their associations with environmental gradients. Hydrobiologia 529: 83-97

Brosse, S., Giraudel, J.L., and Lek, S. (2001) Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages, Ecological Modelling, 146: 159-166, 2001.

Chirart,J. (2003). Development of a macroinvertebrate Index of Biological Integrity (MIBI) for rivers and streams of the St. Croix River basin in Minnesota. Minnesota Pollution Control Agency . Biological Monitoring Program. St. Paul, MN.
Available at: http://www.pca.state.mn.us/publications/reports/biomonitoring-mibi-stcroix.pdf

Fernández, O.A., Murphy, K.J., López Cazorla, A., Sabbatini, M.R., Lazzari, M.A., Domaniewski, J.C.J., Irigoyen, J.H. (1998). Interrelationships of fish and channel environmental conditions with aquatic macrophytes in an Argentine irrigation system. Hydrobiologia 380:15-25

Genet, J., Chirart, J.(2004).). Development of a macroinvertebrate Index of biological Integrity (MIBI) for rivers and streams of the Upper Mississipi river basin. Minnesota Pollution Control Agency.Biological Monitoring Program. St. Paul, MN.
Available at: http://www.pca.state.mn.us/publications/reports/biomonitoring-mibi-uppermiss.pdf

Hall Jr., L.W., Morgan II, R.P., Perry, E.S., Waltz, A. (1999). Development of a Provisional Physical Habitat Index for Maryland Freshwater Streams. Maryland Department of Natural Resources. Chesapeake Bay and Watershed Programs. Monitoring and Non-tidal Assessments

Hilsenhoff, W.L.(1987). An improved biotic index of organic stream pollution. Great lakes Entomol. 20:1-39

Jongman R.H.G., Ter Braak, C.J.F., Van Tongeren,O.F.R. (1995). Data Analysis in Community and Landscape Ecology. Cambridge University Press

Karr, J.R., Chu, E.W. (2000). Sustaining Living Rivers. Hydrobiologia 422/423 (Dev. Hydrobiol.149):1-14

Karr, J.R., Fausch, K.D., Angermeier, P.L., Yant, P.R., Schlosser, I.J. (1986). Assessing Biological Integrity of Running Waters. A Method and its Rationale. Spec. Publ. 5, Illinois Natural History Survey, Champaign, IL.

Karr, J.R., Kerans, B.L. (1991). Components of Biological Integrity: Their Definition and Use in Development of an Invertebrate IBI. 1991 Midwest Pollution Control Biologists Meeting. Environmental Indicators: Measurement and Assessment Endpoints. Lincolnwood, Illinois

Kauffman, P.R., Levine,P., Robison, E.G., Seeliger,C., Peck, D.V. (1999). Quantifying Physical Habitat in Wadeable Streams. EPA/620/R-99/003. U.S. Environmental Protection Agency, Washington D.C.

Kohonen, T.(2001) Self-Organizing MapS. Third Edition. Springer, Berlin.

Lazorchak, J.M., Hill, B.H., Averill, D.K., D.V. Peck and D.J. Klemm (editors). (2000). Environmental Monitoring and Assessment Program-Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Non-Wadeable Rivers and Streams. U.S. Environmental Protection Agency, Cincinnati, Ohio

Makarenkov and Legendre (2002). Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology 83(4): 1146-1161

Maryland Biological Stream Survey (1999). Guide to Using 1995-1997 Maryland Biological Stream Survey Data. Available at: http://www.dnr.state.md.us/streams/mbss/mbss_pubs.html

Mason,C.F.(1991). Biology of fresh water pollution, 2nd ed., Longman Science and Technical, J Wiley, NY

Niemela S.N., Feist, M.D. (2002). Index of Biotic Integrity Guidance for Coolwater Rivers and Streams of the Upper Mississipi River Basin. Minnesota Pollution Control Agency . Biological Monitoring Program. St. Paul, MN.
Available at: http://www.pca.state.mn.us/water/biomonitoring/bio-streams-fish.html

Niemela S.N., Feist, M.D. (2000). Index of Biotic Integrity Guidance for Coolwater Rivers and Streams of the St. Croix River Basin. Minnesota Pollution Control Agency Biological Monitoring Progra,. St. Paul, MN.
Available at: http://www.pca.state.mn.us/water/biomonitoring/bio-streams-fish.html

Norris, R.H., Hawkins, C.P. (2000). Monitoring River Health. Hydrobiologia 435: 5-17

Novotny, V. (2003). Water Quality. Diffuse Pollution and Watershed Management. 2nd edition. John Wiley and Sons. NY

Novotny, V., Bartošová, A., O'Reilly, N., Ehlinger,T. (2005). Unlocking the Relationship of Biotic Integrity of Impaired Waters to Anthropogenic Stresses. Water Research 39(2005):184-198

Ohio Environmental Protection Agency (1987). Biological Criteria for the Protection of Aquatic Life. Columbus, OH

Ohio Environmental protection Agency (1989). Biological Criteria for the Protection of Aquatic Life: Volume III: Standardized Biological Field Sampling and Laboratory Methods for Assessing Fish and Macroinvertebrate Communities. Ecological Assessment Section. Division of Water Quality Planning & Assessment

Paul, M.J., Stribling, J.B., Klauda, R.J., Kazyak, P.F., Southerland, M.T., Roth, N.E. (2003). A Physical Habitat Index for Freshwater Wadeable Streams in Maryland. Final Report. Maryland Department of Natural Resources. Monitoring and non-tidal assessment. Available at: http://www.dnr.state.md.us/streams/mbss/mbss_pubs.html

Rankin, E.T. (1989). The Qualitative Habitat Evaluation Index (QHEI): Rationale, Methods and Application. Ohio EPA. Division of Surface Water. Ecological Assessment Section.

Roth, N.E., Southerland,M.T., Chaillou,J.C., Kazyak,P.F., Stranko,S.A. (2000). Refinement and validation of a fish index of biotic integrity for Maryland streams. Maryland Department of Natural Resources. Chesapeake Bay and watershed programs monitoring and non-tidal assessment. Available at: http://www.dnr.state.md.us/streams/pubs/ea00-2_fibi.pdf

Simonson, T.D., Lyons, J., Kanehl, P.D. (1994). Guidelines for Evaluating Fish Habitat in Wisconsin Streams. Gen. Tech. Rep. NC-164. St. Paul, MN. Department of Agriculture, Forest Service. North Central Experiment Station.

Southerland, M.T., Rogers,G.M., Kline,M.J., Morgan, R.P., Boward, D.M., Kazyak,P.F., Klauda, R.J., Stranko, S.A. (2005). New biological indicators to better assess the condition of Maryland streams. Maryland Department of Natural Resources. Chesapeake Bay and watershed programs monitoring and non-tidal assessment. Available at: http://www.dnr.state.md.us/streams/pubs/ea00-2_fibi.pdf

Southerland, M.T., Stribling, J.B. (1995). Status of Biological Criteria Development and Implementation. In Davis, W.W. & T.P. Simon (Eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida: 81-96.

Stribling, J.B., Jessup,B.K., White, J.S. (1998). Development of a benthic Index of Biotic Integrity for Maryland streams. . Maryland Department of Natural Resources. Chesapeake Bay and watershed programs monitoring and non-tidal assessment. Available at: http://www.dnr.state.md.us/streams/pubs/1998_benthic_ibi.pdf

Ter Braak,C.J.F. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology, Vol.67(5): 1167-1179

United States Environmental protection Agency (1998). Development of an Index of Biotic Integrity Expectations for the Lake Agassiz Plain Ecorregion. EPA-905-R-96-005. Chicago, IL. Available at: http://www.pca.state.mn.us/water/biomonitoring/bio-streams-fish.html

Virani, H, Manolakos, E., Novotny, V. (2005). Self Organizing Feature Maps Combined with Ecological Ordination Techniques for Effective Watershed Management. Technical Report No. 4. Center for Urban Environmental Studies. Northeastern University, Boston. Available at: http://www.coe.neu.edu/environment/publications.htm .