Partial Least Square Analyses of Landscape and Surface Water Biota Associations in the Savannah River Basin

Maliha S. Nash*
US EPA, 944 East Harmon, Las Vegas NV 89119.
E-mail: nash.maliha@epa.gov

and

Deborah J. Chaloud
US EPA, 944 East Harmon, Las Vegas NV 89119.
E-mail: chaloud.deborah@epa.gov

Abstract: Ecologists are often faced with problem of small sample size, correlated and large number of predictors, and high noise-to-signal relationships. This necessitates excluding important variables from the model when applying standard multiple or multivariate regression analyses. In this paper, we present the results of applying partial least square (PLS) regression to explore relationships among biotic indicators of surface water quality and landscape conditions accounting for the above problems. Available field sampling and remotely sensed data sets for the Savannah Basin are used. We were able to develop models and compare results for the whole basin and for each ecoregion (Blue Ridge, Piedmont, and Coastal Plain) in spite of the data constraints. The amount of variability in surface water biota explained by each model reflects the scale, spatial location and the composition of contributing landscape metrics. The landscape-biota model developed for the whole basin using PLS explains 43% and 80% of the variation in water biota and landscape data sets, respectively. Models developed for each of the three ecoregions indicates dominance of landscape variables which reflect the geophysical characteristics of that ecoregion.

Key words: PLS; landscape ecology; water quality; macroinvertebrate; Savannah River Basin.

1. Introduction

The primary objective of the U.S. Environmental Protection Agency's (EPA) Landscape Ecology research program is investigation of associations among indicators of water quality and landscapes. Statistically valid predictive models are an important means of expressing these associations. The analyses presented here represent an attempt to develop a statistical predictive model of biotic indicators of water quality based on associations with a selected suite of landscape indicators.

Investigation of associations among indicators of water quality and landscapes involves statistical analyses of fundamentally different data sets. Data on surface water conditions are generally obtained through field sampling programs and may include several different methods of data production, i.e., on-site observation, chemical analysis of collected samples, and expert identification of biotic organisms. Each method is unique in its precision and variability. Field samples are representative of specific points or stream reaches. Field/analysis programs are expensive and labor intensive; consequently, the total number of sample sites is usually small. The data base may contain missing values due to the realities of field sampling: malfunctioning equipment, lost or destroyed samples, invalidation of results due to poor quality control. Much of the data on watershed characteristics, or landscape data, is derived from remote sensing platforms, thereby permitting wall-to-wall coverage, although the data may be of lesser or more questionable quality than surface water sample data. The landscape indicators data sets may contain a very large number of variables, although many of these are not wholly independent (i.e., they may be collinear).

The characteristics of these data sets and the differences between point sample collection and remote sensing derivation present a challenge in selection of statistical methods for data

analyses and model definition. Single- and multiple-regression analysis has frequently been used to relate water nutrient concentrations to selected landscape variables [1, 2, 3]. Regression analyses, however, are sensitive to missing values and dependence of explanatory variables (landscape variables). Reliable statistically significant results generally cannot be obtained unless the total number of samples greatly exceeds the number of variables. Canonical correlation is well suited to exploring the relationships among two or more distinct data sets to describe their association and connection to the physical environment [4, 5, 6]. However, canonical correlation is sensitive to collinearity in predictors and requires multinormal data sets when testing the significance level of the correlation. The ratio of the number of variables to sample size is critical in canonical correlation; a ratio of 0.025 – 0.05 at a minimum is recommended [7, 8, 9].

Partial least squares (PLS) analysis offers a number of advantages over the more traditionally used regression analyses. Through its extensive use in the field of chemometrics, PLS has been shown to produce significant results when the number of samples is small compared to the number of variables [10, 11, 12]. It has been found to be useful both for providing accurate predictions and for interpreting relationships between data sets containing a high degree of collinearity [13, 14, 15]. Additionally, the prediction error in PLS is smaller than in other multivariate methods [16, 17]. Although PLS is a primary statistical tool in chemometric studies, it has only occasionally been used in ecological studies for exploratory analyses in engineered revegetation studies [17, 18].

The advantages of PLS, described above, makes it an attractive candidate statistical tool for development of landscape ecology models. In this paper, we present the results of applying PLS to exploration of the relationships among surface water biota and landscape conditions.

Available real-world data sets for the Savannah Basin and its three component ecoregions are used. These data sets contain all of the limitations that hinder use of other multivariate statistics, i.e., small number of sampling sites and large number of variables.

2. Methods and Materials

The water data used in this analysis were provided by EPA Region IV, Science and Ecosystem Support Division. As a Regional Environmental Monitoring and Assessment Program (REMAP) project, site selection and sampling were completed according to standard EMAP protocols. This included a random site selection process, with wadeable stream (generally first to third Strahler order) sites selected without consideration of watershed size, proximity to other sampling sites, ecoregion, or ease of access. Sample collection was completed one time during base flow conditions (generally late summer into fall), although selected sites may be visited a second time for quality assurance purposes. Site coordinates were checked with a global positioning system (GPS) unit and against topographic maps to verify the selected sampling location. Macroinvertebrate samples were collected over a 100-m stream stretch above the water sampling point and, at some sites, fish samples were also collected. Macroinvertebrate identification was completed in a biological laboratory following collection. Stream water samples were collected and filtered for subsequent laboratory analyses. All collected samples were sealed, labeled, and transported in coolers under chain-of-custody [19].

For each of the selected sites, the watershed support area was delineated and a suite of landscape variables was calculated [20]. Water biology and landscape variables (n = 86 sites) used in the analyses are described in Table 1. Table 1 also provides the abbreviations of variable names used in the figures and tables in this paper. Due to the great number of variables and the

need for very short abbreviations for use in labeling figures, at each occurrence of a variable name throughout this text, the full variable name will be used with the abbreviation provided in parentheses.

2.1. *Site Description*

The Multi-Resolution Landscape Characterization Consortium (MRLC) landcover/land use data (January 17, 2007; http://www.epa.gov/nerlesd1/land-sci/savannah.htm) reveals distinctive spatial patterns within the Savannah River Basin. The headwaters of the Savannah River are located in the Blue Ridge mountains in which evergreen forests predominate. Below this lies a region of mixed and deciduous forest, agriculture dominated by pasture and hay fields, and several urban centers. Two large reservoirs are located on the main stem river. Below Augusta, Georgia, extensive row crop agriculture is evident, along with wetland areas. The city of Savannah is located near the outlet of the river to the Atlantic Ocean. The spatial patterns seen in the landcover correspond closely to the three ecoregions: Blue Ridge, Piedmont, and Coastal Plain. Two data sets were used (see Table1): four variables for water biology and 26 variables for landscape condition.

Table 1. Water Biota (Response) Variables

Abbreviation	Full name	Methodology
HAB	Macroinvertebrate habitat	A weighted composite score derived from a parameter matrix of on-site observations [21] and modified to fit specific geographical area [19]. Parameters within the matrix are categorized as primary (microscale), secondary (macroscale), or tertiary (riparian zone). Higher scores indicate better conditions for sustaining healthy macoinvertebrate populations.
EPT	Ephemeroptera-Plecoptera-Trichoptera Index	An index of three macroinvertebrate orders known to be sensitive to environmental impacts: Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies), calculated as a percentage of the number of organisms contained in a 100-organism randomly selected subset of the sample collected for macroinvertebrate species richness [21]. In this data set, values >10% indicate non-impacted conditions (>10%) and values $\leq$ 1% indicate severely impacted conditions.

Abbreviation	Full name	Methodology
RICH	Macroinvertebrate Species Richness	A count of the total number of taxa in a sample collected over a 100-m stream reach [21]. Higher numbers indicate a greater diversity of taxa; in this data set, counts > 26 indicated non-impaired conditions and count < 11 indicated severely impacted conditions.
AGPT	Algal Growth Potential Test	Indicator of the amount of nutrients biologically available to support algal growth. A bioassay is performed in the laboratory on aliquots of filtered water collected from the site using standard methodology [22] by Schultz [23]. As a surrogate measurement of nutrient concentration, higher values indicate higher levels of nutrients.

2.2. *Water Biology Variables*

The four water biology variables (Table 1) used in this analysis were Algal Growth Potential Test (AGPT), macroinvertebrate habitat (HAB), macroinvertebrate species richness (RICH), and *Ephemeroptera/Plecoptera/Trichoptera* (EPT).

2.2.1. *Algal Growth Potential Test*

The Algal Growth Potential Test (AGPT) is a bioassay performed in the laboratory in which known amounts of nutrients (nitrogen and phosphorus) and a standard test alga are added to aliquots of filtered water collected from the site [22]. Its purpose is to provide an indication of the amount of nutrients biologically available to support algal growth, as opposed to analytical methodologies that measure the total amount of specific nutrients of which only a portion may be biologically available. The specific methodology used by EPA Region IV was based on the standard method but included a modification by [23] to speed the analytical process [19]. As a surrogate measurement of nutrient concentration, higher values indicate higher levels of nutrients.

2.2.2. *Macroinvertebrate Habitat*

Based on the Rapid Bioassessment Protocols [21] and modified by EPA Region IV to fit their specific ecoregions [19], the macroinvertebrate habitat (HAB) data was derived from visual observations at the sampling site of specific parameters categorized as primary, secondary, and tertiary parameters. Primary parameters characterize the stream habitat at a microscale; these parameters were bottom substrate, available cover, embeddedness, and flow regime. Secondary

parameters characterize stream habitat at the macroscale; these parameters were channel alteration, bottom scouring/deposition, and sinuousity. The tertiary parameters of bank stability, bank vegetation, and streamside cover characterize the riparian zone composition and integrity [19]. From this parameter matrix, a single, weighted composition score was derived, with higher scores indicating better conditions for sustaining healthy macoinvertebrate populations.

2.2.3. *Macroinvertebrate Species Richness*

Macroinvertebrate species richness (RICH) is simply a count of the number of distinct taxa observed in a sample [21]. In this study, samples were collected from a 100-m stream segment above the water sample collection site. D-frame and A-frame dipnets were used to collect organisms from all substrate types within the stream reach [19]. Higher numbers indicate a greater diversity of taxa. The authors assigned ranges based on natural breaks in the data set: non-impacted (greater than 26 taxa), slightly impacted (19 – 26 taxa), moderately impacted (11 – 18 taxa), and severely impacted (less than 11 taxa).

2.2.4. *Ephemeroptera/Plecoptera/Trichoptera*

The *Ephemeroptera/Plecoptera/Trichoptera* (EPT) variable is an index of three macroinvertebrate orders known to be sensitive to environmental impacts: *Ephemeroptera* (mayflies), *Plecoptera* (stoneflies), and *Trichoptera* (caddisflies). It is calculated as a percentage of the number of organisms in these three orders contained in a 100-organism sample [21]. The 100-organism samples used were a randomly selected subset of the sample collected for macroinvertebrate species richness, above. As with macroinvertebrate species richness, the authors assigned

classifications based on natural breaks in the data set: non-impacted (greater than 10 percent), slightly impacted (6 – 10 percent), moderately impacted (2 – 5 percent), and severely impacted (less than 2 percent).

2.3. Landscape Variables

All of the landscape variables used in this analysis were derived from available digital data sets in a geographic information system (GIS). The spatial data sets used were obtained from a variety of sources. The abbreviation, full name and description of each of the landscape variables are given in Table 2. The primary data sets used to derive the 26 variables used in this analysis were: Multi Resolution Land Characteristics (MRLC) Interagency Consortium landcover/landuse [24], State Soil Geographic data base (STATSGO) soils [25], RF3 streams [26], USGS 8-digit HUCs, Georgia and South Carolina subbasins, Region IV sampling site locational data, 30-m and 100-m digital elevation models (DEM) [27], and digital line graph (DLG) roads [28]. Slope was derived as percent rise from the 30-m DEM. Most of the landscape variables were calculated using the derived watershed above the sampling point as the base unit. The single exception in the variables used here is total roads located within 30 meters of a stream (r); for this variable, the base unit was the streams within the watershed, buffered out 30 meters on both sides.

The seven landcover variables were calculated from the MRLC cover classes: Percent crops (c) is the amount of landcover within each watershed identified in the MRLC data as “row crops”, percent pasture (p) is the amount of landcover within each watershed identified in the MRLC data as “pasture or grassland”, percent barren (b) is the amount of landcover within each

Table 2. Landscape (Predictor) Variables

	Full name	Description
c	Percent crop	Percentage of total Multi Resolution Landscape Characterization (MRLC) landcover in row crops types
p	Percent pasture	Percentage of total MRLC landcover in pasture/grassland types
b	Percent barren	Percentage of total MRLC landcover in barren types (Quarries, Strip Mines)
u	Percent urban	Percentage of total MRLC landcover in urban types (Commercial, High- and Low-Density Residential)
f	Percent forest	Percentage of total MRLC landcover in forest types
q	Percent wetlands	Percentage of total MRLC landcover in wetland types
w	Percent water	Percentage of total MRLC landcover in water types
ah	Agriculture on highly erodible soils	Percent of total area in agriculture (row crops + pasture) on highly erodible soils (STATSGO K-factor 0.4)

	Full name	Description
az	Agriculture on slopes >3%	Percent of total area in agriculture (row crops+pasture) on slopes greater than 3 percent
azh	Agriculture on slopes > 3% with highly erodible soils	Percent of total area in agriculture (row crops + pasture) on slopes greater than 3 percent with highly erodible soils (STATSGO K-factor≥ 0.4)
am	Agriculture on moderately erodible soils	Percent of total area in agriculture (row crops + pasture) on moderately erodible soils (STATSGO K-factor≥ 0.2 and < 0.4)
azm	Agriculture on slopes > 3% with moderately erodible soils	Percent of total area in agriculture (row crops + pasture) on slopes greater than 3 percent with moderately erodible soils (STATSGO K-factor≥ 0.2 and < 0.4)
bzh	Barren on slopes > 3% and highly erodible soils	Percent of total area in barren cover types on slopes greater than 3 percent with highly erodible soils (STATSGO K-factor≥ 0.4)
bzm	Barren on slopes > 3% with moderately erodible soils	Barren on slopes > 3% with moderately erodible soils
cz	Crops on slopes > 3%	Percent of total area in row crops on slopes greater than 3 percent

	Full name	Description
czm	Crops on slopes > 3% with moderately erodible soils	Percent of total area in row crops on slopes greater than 3 percent with moderately erodible soils (STATSGO K-factor 0.2 and < 0.4)
pz	Pasture on slopes > 3%	Hay pasture on slope greater than 3 percent
e	Erodible soils	Percent of total area with highly erodible soils (STATSCO K-factor 0.4)
z	Slope > 3%	Percent of total area with slope greater than 3 percent
x	Mean slope	Mean or average percent slope
s	Standard deviation slope	Standard deviation of percent slope
zm	Moderately erodible soils on slopes > 3%	Percent of total area with moderately erodible soils (STATSGO K-factor 0.2 and < 0.4) and slope greater than 3 percent
d	Stream Density	Stream density as total length of streams from USGS TIGER data divided by watershed area

	Full name	Description
v	Total road length within 30 meters of streams	Total length of types 0 through 4 roads and railroads/sidings within 30 m of streams from USGS TIGER data divided by total stream length
r	Total road length in watershed	Total length of types 0 through 4 roads from USGS TIGER data divided by watershed area
t	Total Power, Pipe, and Telephone line length in watershed	Total length of power, pipe, and telephone lines from USGS TIGER data divided by watershed area

watershed identified in the MRLC data as barren due to anthropogenic activities (e.g., “quarries, strip mines”), percent urban (u) is the total amount of landcover within each watershed identified in the MRLC data as “commercial”, “high-density residential”, and “low-density residential”, percent forest (f) is the total amount of landcover within each watershed identified in the MRLC data as “evergreen”, “deciduous”, and “mixed” forest, percent wetlands (q) is the total amount of landcover within each watershed identified in the MRLC data as “woody” and “herbaceous” wetlands, and percent water (w) is the amount of landcover within each watershed identified in the MRLC data as “water.”

Slopes (z) were considered to be all areas with greater than 3 percent rise slope, while mean slope (x) is the arithmetic mean of the 30-m slope pixels within the watershed and standard deviation of slope (s) is the first standard deviation of the total number of slope pixels within the watershed. The STATSGO K-factor was used to provide an estimation of slope erodibility; a K-factor greater than or equal to 0.4 was considered “highly erodible” (e) while a K-factor of greater than or equal to 0.2 but less than 0.4 was considered “moderately erodible.” Moderately erodible soils were used in overlays with landcover and slope data, but not as a variable by itself.

Stream density (d) was calculated as the total length of RF3 stream vectors within the watershed divided by the total area of the watershed. Powerlines, pipelines, and telephone lines (t) was calculated as the total length of these vectors from USGS TIGER files within the watershed divided by the total area of the watershed. Total roads within the watershed (r) is the total length of all USGS TIGER file road classes divided by the total watershed area and total roads within 30 meters of streams (v) is that subset of roads located within the buffered stream boundary, divided by the total stream length. The total roads within 30 meters of streams (v)

also included railroads and sidings as these could produce an impact to streams equal to or greater than some passenger vehicle road classes. Railroads, however, were not included in the total roads within the watershed (r) variable.

The remaining eleven landscape variables were overlays of two or three of the landcover, slope, and soil erodibility variables. Five used total agriculture (MRLC data classified as "row crops" and "pasture or grassland") in combination with slope [agriculture on slopes greater than 3 percent, (az)], or soils [agriculture on highly erodible soils (ah) and agriculture on moderately erodible soils (am)], or both [agriculture on highly erodible soils on slopes greater than 3 percent (azh) and agriculture on moderately erodible soils on slopes greater than 3 percent (azm)]. The subclassifications of agriculture overlayed with slopes and/or soils yielded another three variables [row crops on slopes greater than 3 percent (cz), row crops on slopes greater than 3 percent with moderately erodible soils (czm), and pastures or grasslands on slopes greater than 3 percent (pz)]. Landcover classified as barren due to anthropogenic activities overlayed with slopes and soils accounted for two variables [barren on slopes with highly erodible soils (bzh) and barren on slopes with moderately erodible soils (bzm)]. The last overlay variable used was slopes with moderately erodible soils (zm). Other possible overlay variables (e.g., row crops on slopes with highly erodible soils) were not used in this analysis primarily because they were non-existent in the majority of the watersheds. For clarifications, landscape variables abbreviations in Tables 2 and 3 were used in figures.

3. STATISTICAL METHODOLOGY

The PLS method is based on first computing a few relevant projections (latent variables), i.e. linear combinations of the independent or predictor variable X and then using these new variables in a regression equation for predicting the response Y. In contrast, principal components analysis (PCA) uses only the predictors (X). In PLS, both X- and Y- matrices are decomposed into scores- and weights- matrices ($X = TP^T$ where $T^TT = I$ is identity), then Y is estimated as $\hat{Y} = TBV^T$ where B is the regression coefficient and V is linear weight. The matrix column "T" is the latent vectors. Decomposition of the X- and Y- matrices and forming linear combinations continues until the number of latent vectors is equal the number of variables in the X- matrix. PLS begins by:

1- Centering and scaling each of the response (Y) and predictor (X) variables, Y^o and X^o, respectively.

2- Constructing linear combinations of the predictors as: $\delta(score) = X^o\omega(weight)$. Scores are orthogonal.

3- Constructing linear combinations of the responses as: $\mu = Y^o\nu$.

4- Verifying the linear combination in (2) has maximum covariance $(\delta'\mu)$ with the response linear combination in (3); in addition constraints $\omega^T\omega = 1$ and $\delta^T\delta = 1$ should be met.

5- Predicting for both Y^o and X^o by regression on δ (scores):

$$\hat{X}^o = \delta L_x'$$
$$\hat{Y}^o = \delta L_y'$$

where L_x' (= $(\delta'\delta)^{-1}\delta'X^o$) and L_y' (= $(\delta'\delta)^{-1}\delta'Y^o$) are the X- and Y- loadings, respectively.

6- The above steps are for constructing the first PLS factor.

7- Residuals for each X and Y are produced as:

$$X_1 = X^o - \hat{X}^o$$
$$Y_1 = Y^o - \hat{Y}^o$$

8- The second factor is constructed by applying steps 1 through 5 to the residual (7); additional factors are constructed by repeating this process for each residual until the X matrix becomes null.

In interpretation, the scores as well as weights (steps 2 and 3) are computed and plotted in simple scatter plots (Figures 1 and 2). Weights are the contribution of each the predictors in X to the PLS factor. Landscape metrics clustered near the origin indicate these provide little significant contribution to the predictive model. Clusters of variables with approximately equal weights indicate these variables may be collinear. The scores are the regression coefficients of the variables in X and Y regressed upon the various variables in δ and represent how the different manifest variables are related to the scores δ (Figure 2). The scores are sometimes thought of as latent unobservable variables. Detailed discussions of PLS and other methods can be found in [29, 16, 12, 15, 30].

3.1. Validation

Validation of a prediction is always important for assessing the properties of the equation developed. Just testing the model on data already used for building the model is not enough and

can lead to highly overoptimistic results [16]. Cross validation, as used here, was accomplished by dividing the data into five groups, of which one group was left out (test data). The model was fitted on the remaining four groups (training data). The fitted models (n factor model, step 8) were tested via cross validation using the test data sets and the predicted values were compared with that observed to calculate residuals. The sum of squares of these residuals for all models (null- and n- factor models) was calculated giving PRESS (Predictive Residual Sum of Square), which can be used to define the optimum model and, hence, assess the predictive power of the model. A model with number of factors that minimizes PRESS is the optimum one to be chosen. However, several models may have PRESS values that are close and do not differ greatly from the absolute minimum, therefore, it is important to test whether these differences are significant. A statistical test (Hotelling's T^2) was suggested [31] to test the significant differences between root means PRESS of models was used here. The final model was chosen based on the lowest significant PRESS value (Table 3).

3.2. Variable Influence on Projection (VIP)

VIP is also known as variable importance for projection (Wold, 1995). VIP is calculated as:

$$VIP = \sqrt{V * nx}$$

$$V = \sum_{i=1}^{nf} \frac{\omega_{ni}{}^{2} * r_{yi}}{\sum_{i=1}^{nf} r_{yi}}$$

$$\omega_{ni} = \frac{X\omega_i}{USS(X\omega_i)}.$$

$X\omega_i$ is the predictor (here it is landscape variables) weight per each model factor. For example, if the model has three significant factors, then there are three weights for each of the landscape variables. Each weight is normalized (ω_{ni}) by dividing by the uncorrected sum of squares of predictor weights per factor, r_{yi} is the percent variability in the response variables (here, biota data) that is explained by each factor, *nf* is the number of significant factors in the model, and nx is the number of predictor variables. The values of the regression coefficients and the relative importance (VIP) of each predictor can be used to evaluate the contribution of each variable in the PLS model (Figure 3). Regression coefficient values indicate the contribution of each predictor (lines in Figure 3) for an individual response. The VIP value, as indicated in the above equations, is based on both response and predictor measures. Therefore, if the VIP for a predictor is small in value, it implies that variable has a relatively small contribution to the prediction and may be deleted from the PLS model. Variable with VIP values of less than 0.8 should be considered small contributors [15]. An improved model can be built by including variables with high VIP values and excluding others with low VIP. For the whole Basin, we refined the preliminary model by removing 18 predictors which increased the amount of variability explained by the responses by 9% (Table 4) in the refined model.

The quality of the model developed here was determined by examining the residuals for both the biota and the landscape variables. An examination of any possible outliers using residuals and leverages was carried out to finalize the fitted PLS model. The above analyses

Table 3. Root minimum of predictive residual (PRESS) and its statistics, and percent variation accounted for by the three PLS significant factors. Factors for preliminary PLS model for the surface water biota (4) and landscape (26) variables. Only eight factors were shown below. Bold number denotes the first absolute minimum root means PRESS and its statistics.

# Factors	0	1	2	3	4	5	6	7	8
Root Mean PRESS	1.071	0.997	0.984	**0.968**	0.998	1.047	1.132	1.149	1.179
T^2	17.064	12.556	7.709	**0.000**	5.748	12.00	3.580	5.099	3.647
$P > T^2$	0.001	0.006	0.092	**1.000**	0.172	0.003	0.493	0.245	0.468
Variation in Landscape (%)		25.180	19.978	9.727	Total	54.88			
Variation in Biota (%)		21.754	6.676	5.374		33.80			

were done on all data, and by ecoregion to demonstrate the utility of PLS in different geographical settings.

3.3. Predictive Capabilities Usage

Water quality data (response variables) are predominantly collected by manual methods at selected points. Often, permitting restrictions, cost of sampling, equipment malfunction or other reasons may prohibit collection of a complete set of samples. Landscape variables (predictors), on the other hand, can generally be obtained for all sites. Use of satellite imagery provides nearly complete spatial coverage of the data used in computation of landscape variables. A low

Table 4. Number of sites (n), response and predictor variables, the relative importance (VIP) of each predictor, root mean predictive residual sum of squares (PRESS) for models without (null) and with predictors and percent variability explained by responses and predictors.

	n	Response Variables	VIP	Predictor Variables	PRESS* Null	PRESS* Model	% Variability Responses	% Variability Predictors
Whole Basin								
All data	86	AGPT,EPT,HAB,RICH	>1.0	s,x,f,z,az,p,cz	1.071	0.968	34	55
			1.0- 0.8	e,zm,czm,azm,c,pz, ah,azh,u,d,am,r				
			<0.8	q,b,w,bzm,v,t,bzh				
Refined		EPT,HAB	>1.0	s,z	1.067	0.837	43	80
			1.0- 0.8	e,f,am,c,czm,p,u				
			< 0.8					
Ecoregion:								
Blue Ridge	20	EPT, HAB	>1.0	s, r	1.033	0.832	59	94
			1.0- 0.8	zm,z,pz,f				
			< 0.8					

Piedmont	59	AGPT,EPT, HAB	>1.0	s,f	1.055	0.937	42	65
			1.0- 0.8	z,e,azh,ah,am				
			< 0.8	u,q				
Coastal Plain	7	AGPT, EPT, HAB	>1.0	e,am	1.200	1.132	65	86
			1.0- 0.8	r,s,x,az				
			< 0.8	u				

"*" number of significant factors were three except for the Coastal Plain, with two factors.
PRESS = root mean predictive residual sum of square for the null and predictors model.

numbers of sites, collinearity in the landscape variables, missing values in water quality parameters, and low signal to noise ratios in relationships between landscape variables and biological data, can all be overcome in describing relationships, quantifying variability, modeling and prediction using PLS. We used *SAS* [32] for all statistical analyses.

4. RESULTS

The results for all PLS models are summarized in Table 4. Two models are presented for the whole basin: a preliminary model in which all of the response and predictor variables are used and a refined model using a selected subset of variables. Table 4 also presents results of models for each of the three ecoregions.

4.1. Whole basin

In the preliminary model, three factors are significant explaining 34% of the variability in the biota and 55% of the variability in the landscape data sets. Figure 1 is a plot of the landscape and biota scores for the first factor, indicating the strength of the relationship between the response and predictor variables in this factor ($r = 0.64$). Landscape metrics weights among the three significant factors (Figure 2) shows that erodible soil, slope standard deviation, mean slope, agriculture on slopes, and pasture are heavily weighted in all three factors, while forest is heavily weighted only in factor 1, wetlands in factor 2, and crops in factor 3. Agriculture-related variables, including overlays with slopes and soils are approximately equal weights within the PLS factor indicating collinearity.

Table 4 shows the predictor variables grouped by VIP value. Figure 3 depicts the regression coefficients for each response/predictor variable combination with the predictor

variables listed in order of increasing VIP value. Landscape variables with regression coefficients close to zero and VIP < 0.8 indicates little or insignificant associations between these landscape metrics and water biota in this study. Based on these low values for both regression coefficient and VIP, the following landscape variables are excluded from further analyses, including the PLS models for the individual ecoregions: barren on slopes with either highly or moderately erodible soils, water, stream density, transmission lines, and roads near streams. Several of the agriculture/slope/soils – related landscape variables have similar VIP values and factor weights (Figure 2), indicating approximately equal contribution to the model. Only those variables with high values for both VIP and regression coefficient are selected to create a final refined model for the whole basin with the strongest possible predictive capability; the nine predictor values selected are shown with shaded VIP bars in Figure 3.

The refined model (Table 4) has three significant factors with predictive ability that is more than twice that of the preliminary model. The three factors explain 43% and 80% of the variation in the biota (response) and landscape variables (predictors), respectively. The importance of the nine landscape variables is high (VIP $\geq$ 0.8). The agriculture-related variables contribute equally and minimally to the model, with VIPs close to 0.84. All of agriculture-related variables have a negative effect on EPT and HAB. The most significant contributors (VIP $\geq$ 1) are slope standard deviation, slope, forest, and erodible soils. Slope standard deviation was the most important variable (VIP = 1.5) and, along with forest, has a positive effect on EPT and HAB. Urban is also an important contributor, but ranks in between the above two groups;

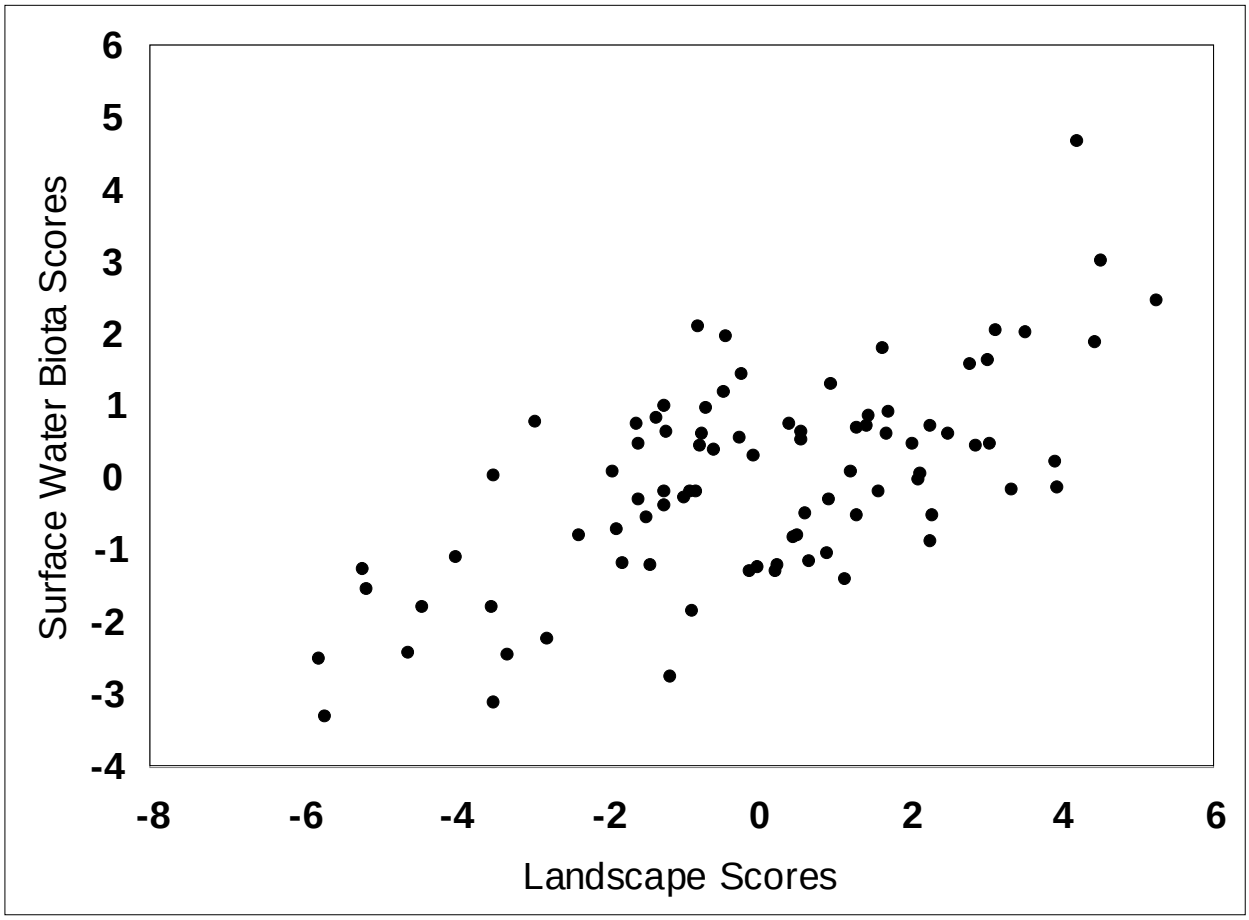

Figure 1. Landscape- and biota- scores for the first PLS factor (correlation is 0.64).

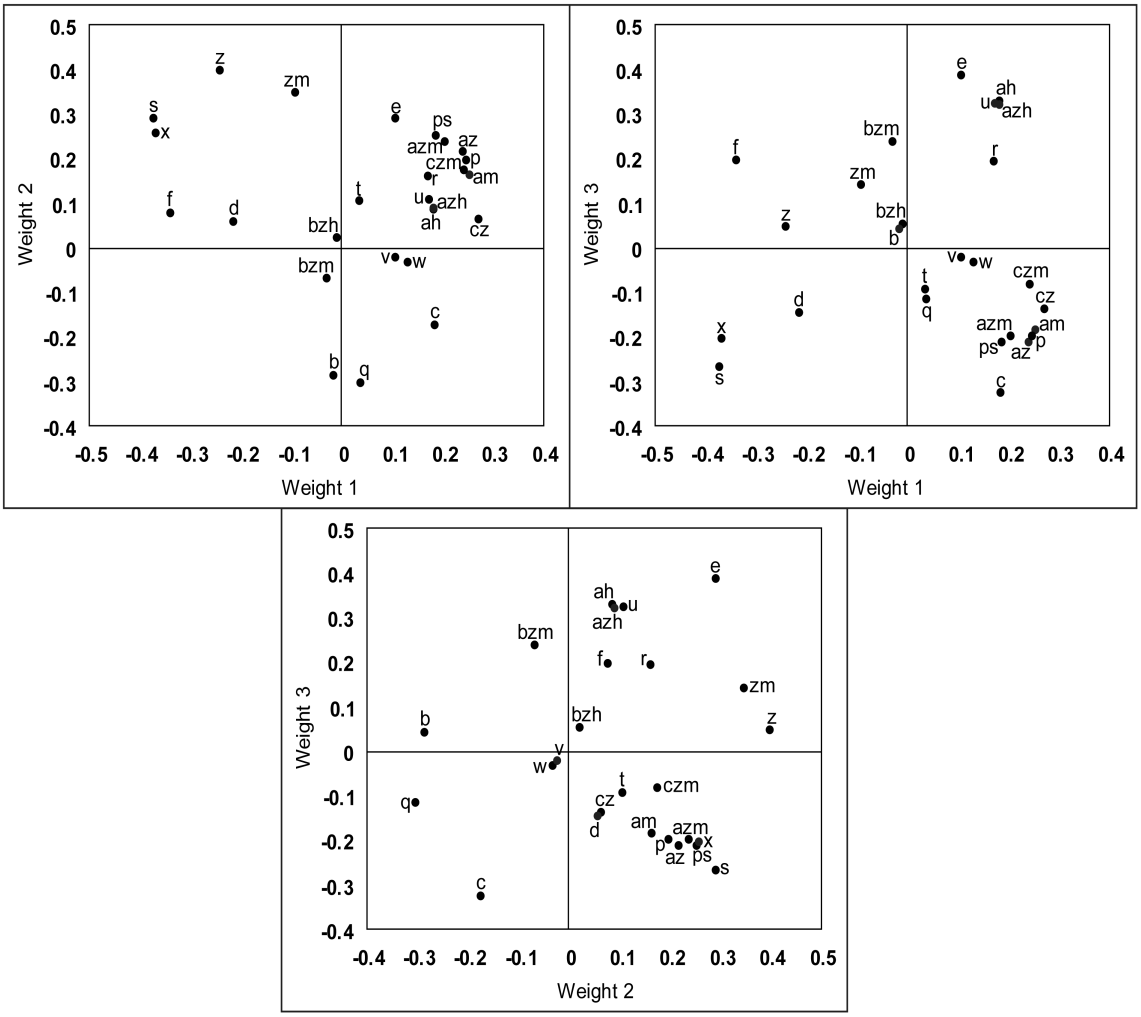

Figure 2. Weight for landscape variables for the three significant PLS factors. See Table 2 for variable description. Those variables that cluster near the origin (i.e., have low weights on both factors) do not contribute much to the predictive capability of the model. Those variables that cluster near each other indicate equal weight on a factor and possibly collinearity.

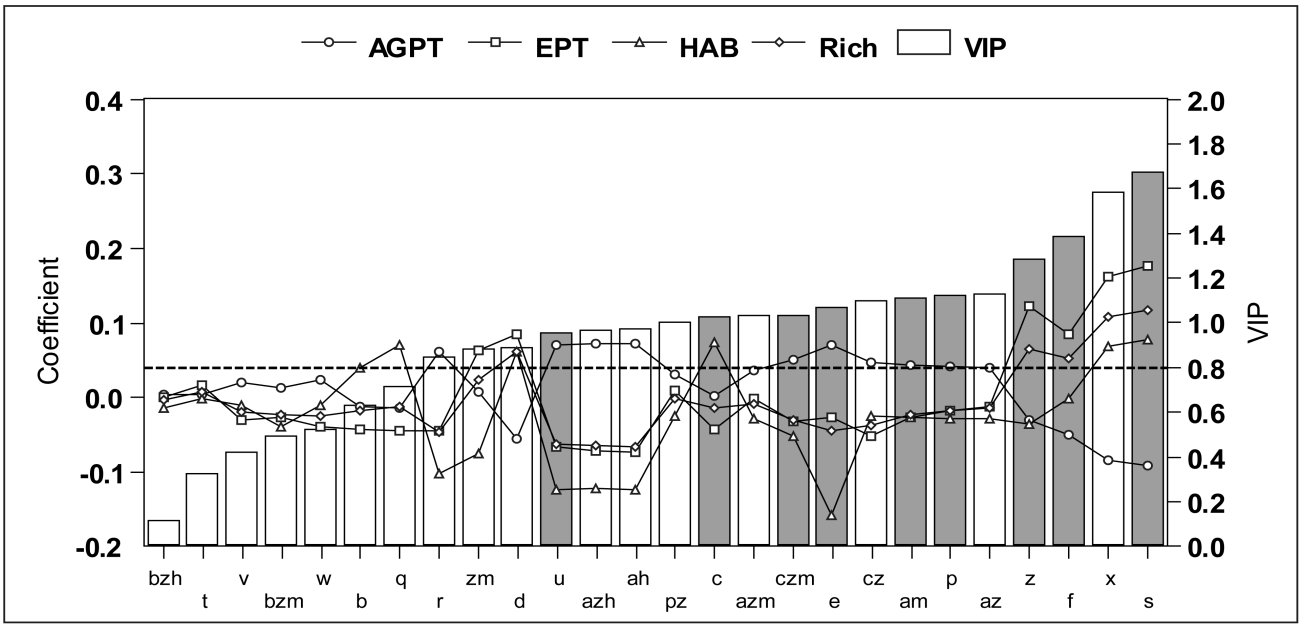

Figure 3. Regression coefficients of each landscape variables on each of response (biota) and the VIP values for each landscape variables in the preliminary model. See Table 2 for variable description. Shaded VIP bars indicate the landscape metrics used in the refined (pruned) model. The regression coefficients were close to zero and their VIP < 0.8 for several landscape variables with respect to biota variables, indicating little or insignificant associations.

i.e., urban contributes more than agriculture, but less than slope standard deviation, forest, and erodible soils. Like agriculture and erodible soils, urban negatively impacts biotic condition.

4.2. Ecoregions

4.2.1 Blue Ridge

Two water biota and six landscape variables (Table 4) for twenty sites are included in the model. There are three significant factors that account for 59% and 94% of the variability for the biota and landscape variables, respectively. The strength of the relationship between the two linear components for the first factor is moderate ($r = 0.65$) [33]. Slope standard deviation and roads are the most important variables (VIP > 1), followed by slopes with erodible soils (VIP ≈ 1; Table 4). Forest, slope and pastures on slopes were ranked in the middle with VIPs greater than 0.8 but less than 1. Slope standard deviation, and to a lesser extent, forest, are positively correlated with EPT and HAB, while the remaining landscape variables are negatively correlated with the biota variables (Figure 4A). Normality of the response is met ($p > 0.13$) and no outliers are found in the landscape metric data.

4.2.2. Piedmont

The PLS model for the Piedmont contains three water biota and nine landscape variables for 59 sites (Table 4), producing three significant factors explaining 42% of the variability in the biota and 65% of the variability in the landscape variables (Table 4). The strength of the relationship between the two linear compositions for the first factor is strong ($r = 0.75$) [33]. Slope features, and forest are the most important variables (VIP > 1; Table 4). Slope features, forest, and

wetlands (marginally) are positively correlated with EPT, whereas agriculture/slope/soils variables, and urban are negatively correlated with EPT (Figure 4B).

Wetlands, slope standard deviation, and forest are positively correlated with HAB while urban and all agriculture-related variables are negatively correlated (Figure 4B). AGPT is heavily weighted and positively correlated with urban and agriculture, and negatively correlated with forest (Figure 4B). Normality of the response variables is met ($p > 0.05$) and no serious outliers in the landscape variables are found.

4.2.3. Coastal Plain

In spite of the scarcity of sampling sites ($n = 7$) in the Coastal Plain, a valid PLS model is constructed. Three water biota variables and seven landscape variables (Table 4) are included in the model. There are two significant factors that account for 66% and 86% of the variability for the biota and landscape variables, respectively. The strength of the relationship between the two linear compositions for the first factor is strong ($r = 0.85$). HAB and AGPT are positively correlated with erodible soils and agriculture on moderately erodible soils, and negatively correlated with the remaining predictor variables. EPT is negatively correlated with agriculture on moderately erodible soils, erodible soils, and urban and positively correlated with the remaining variables (Figure 4C).

5. DISCUSSION

PLS models for the whole basin and for each of the three ecoregions revealed different sets of variables of landscape variables that have relation with that of water quality. A number of variables are found to have little or no contribution to the predictive capability of the model and, therefore, can reasonably be excluded from refined analyses. Variables which are known to have

a high degree of collinearity (specifically, the various overlays of agriculture/slopes/soils) are correctly identified in the analyses with similar weights, VIP values, and regression coefficients. This clustering permits further reduction of the number of variables in refined analyses. From an initial pool of 26 landscape variables, final models are produced with six to nine variables, all significant contributors to the predictive model.

On the ecological aspect, one may ask, do the PLS results identify meaningful associations between biotic and landscape indicators? With the exception of the Coastal Plain, this objective is successful. Macroinvertebrate indicators are positively correlated with natural landcover types (forest in the Blue Ridge and Piedmont, wetlands in the Piedmont) and negatively correlated with indicators of anthropogenic activities (agriculture, urban development, roads). As an indicator of nutrient enrichment, AGPT could be expected to be positively correlated with agriculture and erodible soils. Positive correlations are obtained in the models for the whole basin and for the Piedmont. In the Coastal Plain model, however, AGPT is positively correlated with agriculture on moderately erodible soils and with erodible soils, but is negatively correlated with agriculture on slopes. Also, EPT is negatively correlated with erodible soils and agriculture on moderately erodible soils as could be expected, but EPT is positively correlated with agriculture on slopes and with roads, which is contrary to what would be expected. Although the positive correlation is lower in magnitude than that of the negative, this unexpected relationship is possible due the collinearity between predictors. Soils in this ecoregion are generally of low erodibility and the terrain is much flatter than the other two ecoregions, so possibly the detrimental effects of agricultural runoff are greatly lessened.

The model results also indicate slope is a significant predictor variable in the whole basin and in each of the ecoregions. In the Blue Ridge, slope variables receive the highest weightings

and VIP values. This region is the upland headwaters of the Savannah River Basin and is characterized by hilly to mountainous terrain. Slope variables are also heavily weighted in the Piedmont and standard deviation of slope produces the highest VIP in any of the ecoregion-specific models. The Piedmont is a transitional zone between the mountains of the Blue Ridge and the flat terrain of the Coastal Plain and encompasses terrain varying from hilly to nearly flat. Slope is significant in the Coastal Plain model, but not as much as in the other ecoregions. Unlike the Blue Ridge and Piedmont, standard deviation of slope in the Coastal Plain is negatively correlated with HAB. This may be a function of the methodology used to score HAB which gives higher weights to areas with a variety of pool and riffle habitats. The Coastal Plain may lack this variety due to the lack of slope in this ecoregion.

An unexpected result is seen in the preliminary model for the whole basin. Forest is weighted heavily only on factor 1, wetlands only on factor 2, and row crops only on factor 3 (Figure 2). Forests are the dominant landcover type in the Blue Ridge and row crops are dominant in the Coastal Plain. Wetlands are a small percentage of the total landcover in the Piedmont, but may play a critical role in water quality [20]. It appears the linear combinations in the whole basin model factors may correspond to the characteristics which distinguish individual ecoregions. This result merits further investigation.

Species richness in 362 1-km^2 grid squares in the Kevo Nature Reserve, Finland, were predicted using 227 vascular plant taxa and 27 environmental variables [17]. The resultant PLS model contained two factors which explained 40.3% of the variance in the single response variable. PLS was also used to relate riparian plant growth and survival to duration and frequency of flooding in a controlled experimental study [18]. The availability of remote sensing data for an area can be used to monitor vegetation indicator continuously over time and space

with cost effective and ease of implementation more than that with field measurements. Schmidtlein [34] used transformed reflectance in 64 wavelength bands to predict averaged Ellenberg indicator values (soil pH, soil fertility and water supply) from 46 field sites using PLS regression. In each field site, all vascular plant species were also identified and their cover was estimated. Predicted Ellenberg indicators for the study area were mapped showing the continuous environmental gradient that can be used to assess the floristic composition.

These studies used plant indicators as the response variable and a variety of predictor variables. Our approach using PLS regression differs from these studies in a number of ways: we use multiple response variables, our response variables are indicators of nutrients and macroinvertebrates, and our response data originated from ambient field sampling rather than from controlled experimental studies. These differences are encouraging in that it implies PLS may have utility in a broad range of ecological studies.

6. CONCLUSIONS

In both the preliminary and refined models for the whole basin, associations among water biota and landscape variables largely conform to known ecological processes. Agriculture and urban variables, with their potential for nutrient runoff from fertilizer usage, are positively associated with AGPT measurements while forest is negatively associated with AGPT. Agriculture, urban, moderately eroded soils on slopes, and roads are negatively associated with HAB while wetlands, which filter and remove pollutants as well as slow runoff, are positively associated with HAB.

In each case the dominant landscape variable corresponds to a critical aspect of the ecoregion; forest in the evergreen forest-dominated Blue Ridge, wetland in the transitional

Piedmont, and row crops in the agriculture-dominated Coastal Plain. For both the Blue Ridge and the Coastal Plain, the ecoregion-specific model yields improved results over the basin-wide model, despite the reduction in sample size. Only the Piedmont model fails to improve on the basin-wide model results, with 42% of the variability in the water biota data set and 65% of the variability in the landscape variables explained by three significant factors on a sample size of 59. The Piedmont is a transitional zone with pasture dominant in the upper region transitioning to row crop dominated agriculture in the lower region. Spatial variation across the ecoregion may at least partially explain the model results. In contrast, three significant factors in the Blue Ridge together explain 59% of the variability in the water biota data set and 94% of the variability in the landscape variables data set, based on a sample size of 20. Even with a very limited sample size of 7, the PLS model for the Coastal Plain yields two significant factors, together explaining 66% of the variability in the water biota data and 86% of the variability in the landscapes data.

Although further testing in different biogeophysical setting is needed, the results indicate PLS may prove to be a valuable statistical analysis tool for ecological studies. The data sets used in these analyses contain limitations typical of ecological studies: a small number of sampling sites, a large number of variables, missing values, low signal to noise ratio, differences in spatial extent, and different collection methodologies between the field-collection surface water samples and the remote sensing-derived landscape variables. The PLS methodology is less sensitive to these limitations than other statistical methods. The correlations among water biota variables and landscape variables provide much more information when they are all considered in multivariate regression than in univariate-multiple regressions. Univariate-multiple regression analyses with these data sets will not reveal a distinctive pattern of association due to a weak correlation. Summarizing information in the predictor variables by reduction into a few

variables, i.e. latent variables, conditioned on maximum covariance with the linear composition of the predictor variables, makes PLS more suitable in a multivariate context than other, more commonly used, multivariate methods.

Acknowledgments and Notice

We thank the effort of Dr. Chad Cross, Dr. Tormod Næs, Daniel Heggem and the anonymous reviewers for their input and review. The contribution of the U.S. Environmental Protection Agency, Region IV, Science and Ecosystem Support Division in the collection of the surface water data used in the statistical analyses presented here is gratefully acknowledged. The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and performed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFRENCES

1. I.Noy-Meir; 'Multivariate analysis of the semiarid vegetation in south-eastern Australia. II Vegetation Catenae and environmental gradients', Aust J. Bot, **22**, 115 **(1974)**.

2. .K.Jones, A.C.Neale, M.S.Nash, R.D.Van Remortel, J.D.Wickham, K.H.Riitters, R.V.O'Neil; 'Predicting Nutrient and Sediment Loadings to Streams from Landscape Metrics: A Multiple Watersheds study from the United States Mid-Atlantic Region', Landscape Ecology, **16,** 301 **(2002)**.

3. H.Mehaffey, T.G.Wade, M.S.Nash, C.M.Edmonds; Analysis of Land Cover and Water Quality in the New York Catskill-Delaware Basins, pg. 1327-1339 in D.J.Rapport, W.L.Lasley, D.E.Rolston, N.O.Nielsen, C.O.Qualset, A.B.Damania Eds. 'Managing for Healthy Ecosystems', Lewis Publishers, Boca Raton, Florida (USA) **(2003)**.

4. M.S.Nash, D.J.Chaloud; 'Multivariate Analyses (Canonical Correlation Analysis and Partial Least Square, PLS) to Model and Assess the Association of Landscape Metrics to Surface Water Chemical and Biological Properties using Savannah River Basin Data', Las Vegas (NV): Report no.EPA/600-R-02-091, **(2002)**.

5. S.Cumming, P.Vernier; 'Statistical models of landscape pattern metrics, with applications to regional scale dynamic forest simulations', Landscape Ecology, **17,** 433 **(2002)**.

6. M.S.Nash, D.J.Chaloud, S.E.Franson; 'Association of Landscape Metrics to Surface Water Biology in the Savannah River Basin', Journal of Mathematics and Statistics, **1,** 29, **(2005)**.

7. R.Barcikowski, J.P.Stevens; 'A Monte Carlo study of stability of canonical weights, and canonical variate-variable correlation', Multivariate Behavioral Research, **10,** 353 **(1975)**.

8. R.M.Thorndike; ‘Correlation procedure for research’, Gardner Press, New York. **(1978)**.

9. R.Gittins; ‘Canonical analysis: A review with application in ecology’, Springer Verlag, NewYork **(1985)**.

10. W.Lindberg, P.Jane-Ake, S.Wold; ‘Partial Least-Square method for Spectrofluorimetric analysis of mixture of humic acid and lignisulfonate’, Anal. Chem., **55**, 643 **(1983)**.

11. S.de Jong, H.A.L.Kiers; ‘Principal covariates regression. Part I Theory’. *Chemometric and Intelligent Laboratory Systems*, **14,** 155 **(1992)**.

12. L.E.Frank, J.H.Friedman; ‘A statistical view of some chemometrics regression tools’, Technometrics **35:**109-135 **(1993)**.

13. I.S.Helland; ‘On the structure of partial least square regression’, Commun. Statist. Simula., **17,** 581 **(1988)**.

14. S.Wold; ‘The collinearity problem in linear regression: The partial Leats squares (PLS) approach to generalized inverses’, Soc. Indus. App. Math.: J. Sci. Stat Comput., **5,** 735 **(1984)**.

15. S.Wold; PLS for multivariate Linear Modeling, pg. 195-218, in H.van de Waterbeemd Eds, ‘Chemometric methods in molecular design methods and principles in medicinal chemistry’, Weinheim, Germany: Verlag-Chemie, **(1995).**

16. H.Martens, T.Næs; ‘Multivariate calibration’, John Wiley and Sons, Chichester (England) **(1989)**.

17. R.K.Heikkinen; ‘Predicting patterns of vascular plant species richness with composite variables: A meso-scale study in Finnish Lapland’, Vegetation, **126(2),** 151 **(1996)**.

18. M.E.Johansson, C.Nilsson; 'Responses of riparian plants to flooding in free-flowing and regulated boreal rivers: an experimental study', Journal of Applied Ecology, **39**, 971 **(2002)**.

19. [USEPA] Environmental Protection Agency; A Demonstration of the Usefulness of Probability Sampling for the Purpose of Estimating Ecological Condition in State Monitoring Programs, Athens (GA): Region IV, Report no.EPA/904-R-99-002 **(1999).**

20. D.J.Chaloud, C.M.Edmond, D.T.Heggem; 'Savannah River basin landscape analysis', Las Vegas (NV): U.S. Environmental Protection Agency, Office of Research and Development, Report no.EPA/600-R-01-069 **(2001)**.

21. M.T.Barbour, J.Gerritsen, B.D.Snyder, J.B. Stribling; 'Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates, and Fish', 2nd Eds. Washington (DC): U.S. Environmental Protection Agency, Office of Water, Report no.EPA/841-B-99-002. **(1999)**.

22. APHA; 'Standard Methods for the Examination of Water and Wastewater', 19th Ed., American Public Health Association, Washington, D.C., **(1995)**.

23. D.Schultz, R.Raschke, R.Jones; 'A shortened algal growth potential test'. Environmental Monitoring and Assessment, **32,** 201 **(1994)**.

24. T.J.Bara; 'Multi-Resolution Land Characteristics Consortium' – Documentation Notebook. U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, **(1994)**.

25. Natural Resource Conservation Service; 'State Soil Survey Geographic Data Base (STATSGO) Metadata'. U.S. Department of Agriculture, Natural Resource Conservation Service, Washington (DC) **(1996)**.

26. [USEPA] Environmental Protection Agency; The U.S. EPA reach file version 3.0 Alpha Release (RF3-Alpha), Technical Reference, Washington (DC): Office of Wetlands, Oceans, and watersheds, Office of Water, **(1994).**

27. [USGS] US Geological Survey; 'Digital Elevation Models, National Mapping Program Technical Instructions, Data Users Guide 5', 2nd Printing (Revised), Reston (VA) **(1990)**.

28. [USGS] US Geological Survey; 'Digital Line Graphs from 1:100,000-Scale Maps – Data Users Guide 2', Reston (VA) **(1989)**.

29. T.Næs, H.Martens; 'Comparison of predicted methods for multicolinear data', Commun. Statist. -Simula. Computa., **14,** 545 **(1985)**.

30. T.Hastie, R.Tibshirani, J.Friedman; 'The elements of statistical learning; Data mining, inference, and prediction', Springer, New York, **(2001)**.

31. H.van der Voet; 'Comparing the predictive accuracy of models using a simple randomization test', Chemometric and Intelligent Laboratory Systems, **25,** 313 **(1994)**.

32. SAS; 'Stat User's Guide'. *SAS* Institute. Inc. Cary, NC, USA. **(1998)**

33. D.J.Sheskin; 'Handbook of Parametric and nonparametric statistical procedures', Chapman & Hall/CRC, New York, **(2000).**

34. S.Schmidtlein; 'Imaging spectroscopy as a tool for mapping Ellenberg indicator values', Journal of Applied Ecology, **42,** 966 **(2005)**.