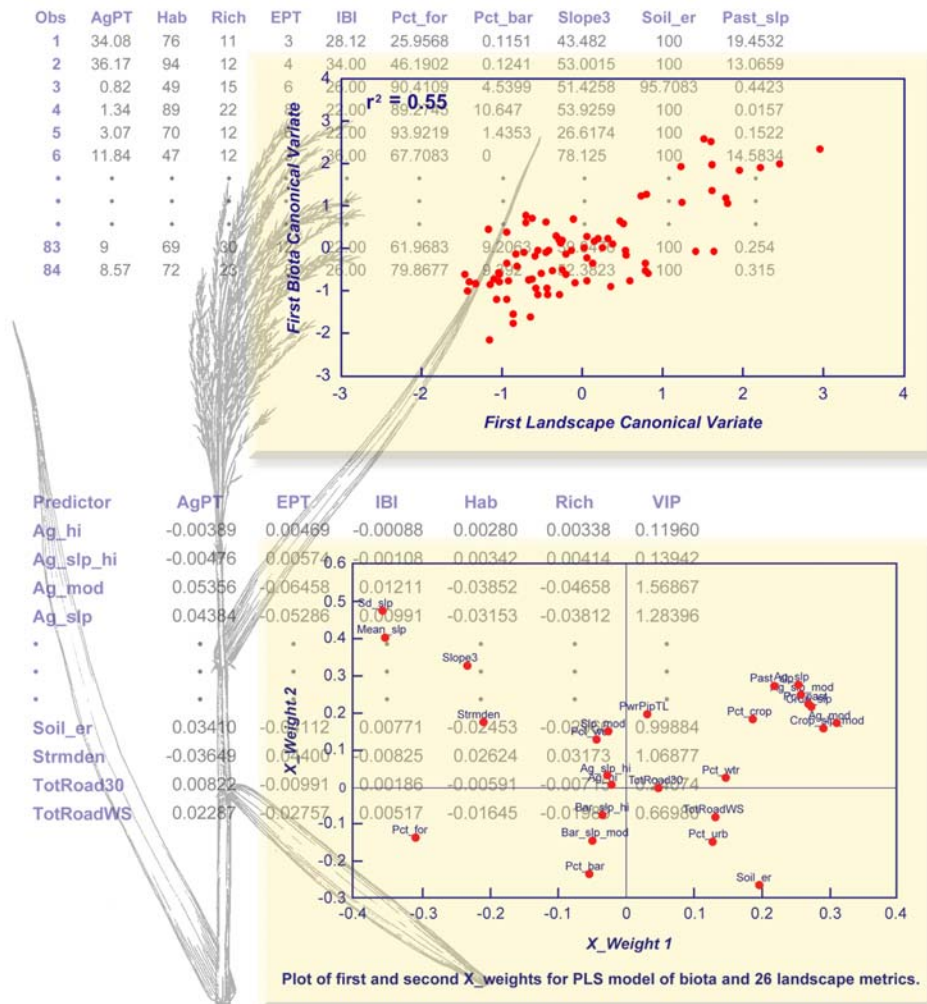


Applications of Canonical Correlation and Partial Least Square Analyses in Landscape Ecology



Applications of Canonical Correlation and Partial Least Square Analyses in Landscape Ecology

by

Maliha S. Nash, Deborah J. Chaloud, and Ricardo D. Lopez

Environmental Sciences Division
U.S. Environmental Protection Agency
Las Vegas, Nevada

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Las Vegas, Nevada 89119

Notice

This is a second edition of a report first titled “Multivariate Analyses (Canonical Correlation 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” by Maliha S. Nash and Deborah J. Chaloud that was published in November 2002 (EPA/600/R-02/091). One reason for this second addition was our desire to apply partial least square analyses to different landscape ecological data. Dr. Ricardo Lopez provided data from a Western Lake Erie wetland in an effort to predict invasive plant (*Phragmites australis*) height and stem density solely using cover measurements. Data were provided from 221 sampling sites.

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.

Acknowledgement

The authors would like to acknowledge the review and valuable input of Dr. Ann Pitchford and Dr. David Bradford, who made this report more comprehensive and easy to follow. Also, we would like to thank Robert Furlow for his input.

Table of Contents

	<u>Page</u>
Notice	iii
Acknowledgement	iii
List of Tables	vii
List of Figures	ix
Section 1 Introduction	1
Section 2 Goals and Objectives	3
Section 3 Data Sets	5
Section 4 Study Site Description	9
Section 5 Statistical Analyses	11
5.1 Canonical Correlation Analysis	11
5.1.1 Canonical Variate and Model Rank	12
5.1.2 Squared Canonical Correlation	12
5.1.3 Canonical Coefficients	13
5.1.4 Correlation Between the Original Variables and Canonical Variates	13
5.1.5 Variables Entering the Canonical Correlation Analysis	15
5.1.6 Results	15
Pairwise Correlation of the Original Variables	15
a) Biological and Landscape Metrics	15
b) Chemical and Landscape Metrics	15
c) Chemical and Biological Metrics	15
Canonical Correlation	15
a) Biological and Landscape Metrics	15
b) Chemical and Landscape Metrics	16
c) Biological and Chemical Metrics	16
Canonical Coefficient	16
a) Biological and Landscape Metrics	20
b) Chemical and Landscape Metrics	20
c) Chemical and Biological Metrics	20

Table of Contents, Continued

	<u>Page</u>
Canonical Structure (Canonical Redundancy Analyses)	21
Within Each Data Set (Intra-set Correlation)	21
a) Biological and Landscape Metrics	21
b) Water Chemistry and Landscape Metrics	21
c) Water Chemistry and Biological Metrics	21
Between Data Sets (Inter-set Correlation)	21
a) Biological and Landscape Metrics	21
b) Water Chemistry and Landscape Metrics	22
c) Water Chemistry and Biological Metrics	22
R ² and Canonical Redundancy Analysis	22
5.1.7 Summary	23
5.2 Partial Least Square (PLS) Analysis	24
5.2.1 Fitting the Model	25
5.2.2 Diagnostic Checking and Variable Influence on Projection (VIP)	25
5.2.3 PLS Analyses	29
5.2.3.1 Savannah River Basin	29
All Data	29
By Ecoregion	33
a) Blue Ridge (BR)	33
b) Piedmont (P)	35
c) Coastal Plain (CP)	37
Prediction for the Non-Sampled Sites	39
Correlation Between the Dependent Variables	40
Summary	43
5.2.3.2 Application of PLS to Other Ecological Data	43
Western Lake Erie Wetland	43
Results	44
Conclusion	48
Section 6 Comparison with Canonical Correlation	49
References	51
Appendix I Supplementary Information	53
Missing Observations	53
Assumptions	53
1 - Collinearity	53
2 - Normality	54
3 - Number of Variables to Number of Observations Ratio	57
Appendix II %multinorm SAS™ Macro	61
Appendix III SAS™ Statements for PLS and Diagnostic Checking	67
Appendix IV SAS™ Statements for Predicting the Nonmeasured Dependent Variables	77

List of Tables

<u>Table #</u>	<u>Description</u>	<u>Page</u>
1	Chemical-biological surface water data and landscape metrics used in the analyses	5
2	Canonical correlation (r_k), canonical r-square (r_k^2), percentage trace (% <i>Trace</i> = $r_k^2 * 100 / \sum r_k^2$), and cumulative trace (%) of total (standardized) sample variance explained by each canonical variate and the probability of exceeding the critical value (F)	12
3	Canonical correlation structure: correlation between the original variables and the canonical variate	14
4	The significant factors for the preliminary PLS model for the surface water biological (5) and landscape metrics (26)	26
5	PLS regression coefficients and VIP values for the preliminary model (5 biota and 26 landscape metrics)	27
6	The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the final model	31
7	Regression coefficient values for the 5 biota and VIP for landscape metrics in the final PLS model	31
8	The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the Blue Ridge (BR) ecoregion (n = 19)	33
9	Regression coefficient and VIP values for biological and landscape metrics values for the Blue Ridge (BR) PLS model	34
10	The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the Piedmont (P) ecoregion (n = 59)	35
11	Regression coefficient and VIP values for the Piedmont (P) PLS model	37
12	Percent variation accounted for by PLS factors for the Coastal Plain (CP) ecoregion	37
13	Regression coefficients of the biota and VIP for landscape metrics for the Coastal Plain (CP) ecoregion	39

List of Tables, *Continued*

<u>Table #</u>	<u>Description</u>	<u>Page</u>
14	The significant factors of the final PLS model for the 4 biota and 12 landscape metrics and percent variation accounted for by PLS factors for the Piedmont (P) ecoregion with no missing data (n = 52)	41
15	Variables used in the analyses with a description of each field-sampled plant structural characteristic	44
16	The significant factors for the PLS model for the numL, ht and 11 plant community variables (%LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5)	45
17	Coefficient values for the 2 responses (ht and numL) and 11 predictors (%LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5) and VIP for the 11 plant metrics for the PLS model	47
18	Rank of the landscape metrics in the PLS model using VIP levels	50
I-1	The best fit model for the surface water quality variables with their significant levels . . .	54
I-2	Pairwise correlation between all variables in the biological and landscape data	55

List of Figures

<u>Figure #</u>	<u>Description</u>	<u>Page</u>
1	The Savannah River Basin ecoregion and MRLC classification (a) and the sample locations and erodibility classes (b)	9
2	The scatter plot of the first two pairs of canonical variates for the landscape (Past_slp, Soil_er, Pct_bar, and Slope3) and surface water biological (AgPT, IBI, Hab, Rich, and EPT) data sets	17
3	The scatter plots of the first two pairs of canonical variates for the landscape metrics (Past_slp, Pct_for, Pct_bar, and Slope3) and surface water chemistry (DO, pH, and EC) data sets	18
4	The scatter plots of the first two pairs of canonical variates for the surface water biota (IBI, Hab, Rich, and EPT) and surface water chemistry (DO, pH, and EC) data sets	19
5	Plot of X_Scores (u) and Y_Scores (v) for factor 1 for each observation from the PLS model of 5 biota and 26 landscape metrics	28
6	X_Scores from the first and second PLS factors for each observation	28
7	Plot of first and second X_Weights for the PLS model of the 5 biota and 26 landscape metrics	30
8	Plot of X_ and Y_Scores for the first factor of the final PLS model of the 5 biota and 14 landscape metrics	32
9	Plot of first and second X_Weights for the first PLS final model of the 5 biota and 14 landscape metrics	32
10	Plot of X_ and Y_Scores for factor 1 for each observation for the PLS model of the 5 biota and 14 landscape variables for the Blue Ridge (BR) ecoregion	34
11	Plot of X_ and Y_Scores for factor 1 for each observation for the PLS model of the 5 biota and 14 landscape variables for the Piedmont (P) ecoregion	36
12	Plot of first and second X_Weights for the PLS model of the 5 biota and 14 landscape metrics for the Piedmont (P) ecoregion	36
13	Plot of X_ and Y_Scores for factor 1 for each observation from PLS model of the 5 biota and 14 landscape variables for the Coastal Plain (CP) ecoregion	38
14	Plot of first and second X_Weights for the PLS model of the 5 biota and 14 landscape metrics for the Coastal Plain (CP) ecoregion	38

List of Figures, *Continued*

<u>Figure #</u>	<u>Description</u>	<u>Page</u>
15	Regression coefficients of landscape metrics for each biota and VIP value for the PLS model for each landscape metrics with no missing biota for the Piedmont (P) ecoregion .	41
16	Loading of each of the biota variables (dependent) on the first two principal components (Prin 1 and Prin 2)	42
17	Scores for principal components (Prin 1 versus Prin 2) for the biota variables showing no cluster pattern in sites	42
18	Linear compositions of the predictors (%LNTC, %LTC, %totNTC, %totNTU, %totTC, %LNTU, and D1-D5) of plant community for the three significant PLS factors	46
19	The relation between <i>P. australis</i> height/number and percent cover scores for the first PLS factor ($R^2 = 0.53$)	47
20	Final PLS model residuals: (above) distance from X_variables to the final PLS model and (below) distance from Y_variables to the final PLS model	48
I-1	The squared distance and chi-square quantile for the landscape and biota variables used for the canonical correlation analyses	57
I-2	The x-distance (a) and y-distance (b) to the model	59

Section 1

Introduction

In ecological studies, often many measurements are taken from many sites in an area to analyze and explore relationships. A number of statistical methods may be used to analyze and explore relationships among variables. Single- and multiple-regression analyses have frequently been used to relate water nutrient concentrations to selected landscape metrics (Noy-Meir, 1974; Jones et al., 2001; Mehaffey et al., 2001). These results quantified relationships and answered questions regarding the status of the landscape in an area. But, when the need is to relate two or more distinct data sets (e.g., landscape metrics and chemical properties of surface water) to describe their association and to explain their connection to their physical environment, multivariate analyses, such as canonical correlation and partial least square (PLS), should be used.

Many multivariate methods are used in describing and predicting relationships; each has its unique usage of categorical and non-categorical data. In multivariate analysis of variance (MANOVA), many response variables (y's) are related to many independent variables that are categorical (classes, levels). For example, relating nitrogen, phosphorous and fecal coliform to presence/absence of urban development, farms, soil types, geological formations, etc., can be represented by the generalized equation (nitrogen + phosphorous + fecal coliform = type of farm, urban development, geology, soil . . .). In analysis of variance (ANOVA), a dependent (response) variable is related to an independent variable that is categorical. For example, determining the response of an ant species to grazing level (severe, medium, and low) in an area can be represented by the generalized equation (ant abundance = grazing levels). In multiple discriminant analysis, the dependent variable (y) is categorical (groups or classes) and related to the independent variables (x's). For example, presence/absence of amphibians in an area relates to many environmental variables and can be represented by the generalized equation (pres/abs = percent bedrock substrate cover + water depth + percent vegetation cover + . . .). In multiple regression, the dependent variable (y) is related to many independent variables (x's). For example, nitrogen loading relates to landscape metrics such as percent forest, percent crops, percent of wetland, and percent of urban development.

In canonical correlation, two sets of variables are related and these variables may or may not be categorical. It is an extension of the multiple regression and it is directly related to principal components-type factor analytic models. In the canonical analysis method, a number of composite associations between sets of multiple dependent and independent variables are performed. Consequently, a number of independent canonical functions that maximize the correlation between the linear composites of sets of dependent and independent variables are developed. The main goal of the canonical correlation analysis is to develop these linear composites (canonical variate) and derive a set of weights for each variate. Thereby, this explains the nature of relationships that exists between the sets of response and predictor variables that are measured by the relative contribution of each variable to the canonical functions (relationships) that exist. The result of applying canonical correlation is a measure of the strength of the relationship between two sets of multiple variables. This measure is expressed as a canonical correlation coefficient (r) between the two sets.

Canonical correlation analysis is used to describe the association between two sets of variables such as the relationship between water biological metrics as dependent variables and landscape metrics as independent variables. Canonical correlation analyses results are used to describe association, including vegetation species and environmental conditions (Ter Braak, 1987; Johnson and Altman, 1996). Also, canonical correlation analysis results are used to describe the physical process that leads to vegetation variation as a response to environmental conditions.

PLS is also a multivariate analysis well known in chemometrics for use in studying the structure pattern among groups of chemicals. It is used in the prediction of chemical form from spectroscopy data, where the norm is several hundred wavelengths and a smaller number of chemical samples (Owen, 1988). PLS is used in quantification of molecular modeling, where a large number of independent variables (>1,000) are normally obtained with respect to the number of samples (10-100). PLS features, as will be seen below, make it applicable in relating landscape metrics with water quality properties.

Canonical correlation requires a relatively large number of observations compared to the number of variables. It is also sensitive to collinearity in independent variables and requires multinormal data sets. Multinormality is a requirement when the test of the significance for the canonical correlations is considered to define the model rank. In ecology, sample size is often small, number of independent variables is large, and the independent variables are frequently correlated. This necessitates excluding important variables from the model. These latter problems are overcome by using PLS. Similar to canonical correlation, PLS outputs many statistics that can be used to describe the variability by the two data sets. But PLS predicts for dependent and independent variables, and it also produces the relative importance values of the independent variables. Relative importance values can be used to decide which independent variable contributes the most to the fitted model.

Section 2

Goals and Objectives

Our primary goals were:

- 1) to determine if there are relationships (dependence) between the two data sets of surface water biota and multiple landscape metrics,
- 2) to determine if the relationships are significant between the landscape metrics and surface water properties,
- 3) to quantify the strength of relationships, and
- 4) to define the key landscape variable(s) that contributes to surface water quality.

Therefore, our objectives were:

- 1) to explore the relationships among chemical and biological surface water properties and landscape data sets,
- 2) to quantify the magnitude of the relationship between the chemical and biological variables with the landscape metrics, and
- 3) to investigate the possibility of using chemical data, which is less expensive and easier to measure, as a surrogate for biological data to examine changes in landscape and related effects in water quality metrics.

Section 3

Data Sets

The surface water data used in this analysis was provided by EPA Region 4, 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. For each of the selected sites, the watershed support area was delineated, and a suite of landscape metrics was calculated (Chaloud, 2001). Three data sets: 1) water chemistry; (2) surface water biota, and (3) landscape metrics were used in the analyses (see Table 1 for variable description). Number of variables entered into the model for canonical correlation were different from PLS, because of normality, missing values, and collinearity issues with the canonical correlation (see Appendix I). Variables used for each model are described in each method below.

Table 1. Chemical-biological surface water data and landscape metrics used in the analyses. Letters “c” and “p” denote variables included in canonical and PLS analyses, respectively.

Variable Type/Name	Full Name	Description
<i>Water Chemistry Metrics</i>		
pH (c)	pH	pH
EC (c)	EC	Electrical Conductivity, micro-siemens per centimeter (μ S/cm)
DO (c)	DO	Dissolved Oxygen
<i>Surface Water Biota</i>		
Hab (c, p)	MI_Habitat	Macroinvertebrate habitat.
EPT (c, p)	MI_EPT	Taxa richness of sensitive insects to pollution; EPT stands for the Ephemeroptera-Plecoptera-Trichoptera Index. These insects are correlated with good water quality (Lenat, 1987) based on 100-organism subsample, nonimpacted (>10), slightly impacted (6-10), moderately impacted (2-5) and severely impacted (0-1).
Rich (c, p)	MI_Richness	Macroinvertebrate richness, species richness (SPP). Total number of species in a sample. Condition of areas is classified as: non-impacted (>26), slightly impacted (19-26), moderately impacted (11-18), severely impacted (<11).
AgPT (c, p)	AgPT	Algal growth Potential Test
IBI (c, p)	fish_IBI	Fish Index of Biotic Integrity (Karr, 1981)
<i>Landscape Metrics</i>		<i>Measurements in a Hydrologic Unit Code (HUC)</i>
Pct_crop (c, p)	Percent crop	Percentage of total Multi-Resolution Land Characteristics (MRLC) landcover in row crops types
Pct_past (c, p)	Percent pasture	Percentage of total MRLC landcover in pasture/grassland types
Pct_wtr (c, p)	Percent water	Percentage of total MRLC landcover in water types

Table 1 (Continued)

Variable Type/Name	Full Name	Description
Landscape Metrics		Measurements in a Hydrologic Unit Code (HUC)
Pct_wet (c, p)	Percent wetlands	Percentage of total MRLC landcover in wetland types
Pct_bar (c, p)	Percent barren	Percentage of total MRLC landcover in barren types (quarries, strip mines)
Pct_urb (c, p)	Percent urban	Percentage of total MRLC landcover in urban types
Pct_for (c, p)	Percent forest	Percentage of total MRLC landcover in forest types
Crop_slp (c, p)	Crops on slopes >3%	Percent of total area in row crops on slopes greater than 3 percent
Crop_slp_mod (c, p)	Crops on slopes >3% and on moderately erodible soils	Percent of total area in row crops on slopes greater than 3 percent with moderately erodible soils (State Soil Geographic (STATSGO) data base K-factor ≥ 0.2 and < 0.4)
Ag_hi (c, p)	Agriculture on highly erodible soils	Percent of total area in agriculture (row crops + pasture) on highly erodible soils (STATSGO K-factor ≥ 0.4)
Ag_slp (c, p)	Agriculture on slopes >3%	Percent of total area in agriculture (row crops + pasture) on slopes greater than 3 percent
Ag_slp_hi (c, p)	Agriculture on slopes >3% and on highly erodible soils	Percent of total area in agriculture (row crops + pasture) on slopes greater than 3 percent and on highly erodible soils (STATSGO K-factor ≥ 0.4)
Ag_mod (c, p)	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)
Ag_slp_mod (c, p)	Agriculture on slopes >3% and on moderately erodible soils	Percent of total area in agriculture (row crops + pasture) on slopes greater than 3 percent and on moderately erodible soils (STATSGO K-factor ≥ 0.2 and < 0.4)
Past_slp (c, p)	Ag_slp - crop_slp	Percent of total area in hay pasture on slopes greater than 3 percent
Bar_slp_hi (c, p)	Barren on slopes >3% and highly erodible soils	Percent of total area in barren cover types on slopes greater than 3 percent and on highly erodible soils (STATSGO K-factor ≥ 0.4)
Bar_slp_mod (c, p)	Barren on slopes >3% and on moderately erodible soils	Percent of total area in barren on slopes >3 percent and on moderately erodible soils
Soil_er (c, p)	Erodible soils	Percent of total area with highly erodible soils (STATSCO K-factor ≥ 0.4)
Mean_slp (c)	Mean slope	Mean or average percent slope
Slope3 (c, p)	Slope >3%	Percent of total area with slopes greater than 3 percent
Slp_mod (c, p)	Moderately erodible soils on slopes >3%	Percent of total area with moderately erodible soils (STATSGO K-factor ≥ 0.2 and < 0.4) and slopes greater than 3 percent
Sd_slp (c)	Standard deviation slope	Standard deviation of percent slope
Strmden (c, p)	Stream density	Stream density as total length of streams from USGS Topologically Integrated Geographic Encoding and Referencing (TIGER) data

Table 1 (Continued)

Variable Type/Name	Full Name	Description
Landscape Metrics (Continued)		Percent area in an HUC
Totroad30 (c, p)	Road30t1 +	Total length of type 1 (interstate) roads within 30 m of streams from USGS TIGER data
	Road30t2 +	Total length of type 2 (interstate) roads within 30 m of streams from USGS TIGER data
	Road30t3 +	Total length of type 3 (interstate) roads within 30 m of streams from USGS TIGER data
	Road30t4 +	Total length of type 4 (interstate) roads within 30 m of streams from USGS TIGER data
	Rail30 +	Total length of railroads within 30 m of streams from USGS TIGER data
	Rr30_siding	Total length of railroad sidings within 30 m of streams from USGS TIGER data
TotroadWS (c, p)	Road_t1 +	Total length of type 1 (interstate) roads from USGS TIGER data
	Road_t2 +	Total length of type 2 (interstate) roads from USGS TIGER data
	Road_t3 +	Total length of type 3 (interstate) roads from USGS TIGER data
	Road_t4 +	Total length of type 4 (interstate) roads from USGS TIGER data
	Road_t0	Total length of type 0 (trails) roads from USGS TIGER data
PwrPipTI (c, p)	Powerline30 +	Total length of power lines within 30 m of streams from USGS TIGER data
	Pipeline30 +	Total length of pipe lines within 30 m of streams from USGS TIGER data
	Telephone30	Total length of telephone lines from USGS TIGER data

Section 4

Study Site Description

The Multi-Resolution Land Characteristics Consortium (MRLC) land cover/land use data for the Savannah River Basin (Figure 1a) reveals distinctive spatial patterns. 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 evergreen and deciduous forest, agriculture dominated by pasture/hay and several urban centers. Two large reservoirs can be seen on the main stem river. Below Augusta, Georgia (the large urban center in the middle), extensive row crop agriculture is evident, along with a wetland area. The city of Savannah can be seen near the outlet of the river to the Atlantic Ocean. The spatial patterns displayed in the land cover correspond closely to the four ecoregions: Blue Ridge (BR), Piedmont (P), Coastal Plain (CP), and Atlantic Coastal Plain (ACP) (Figure 1b). For this report, only three ecoregions (BR, P, and CP) were used.

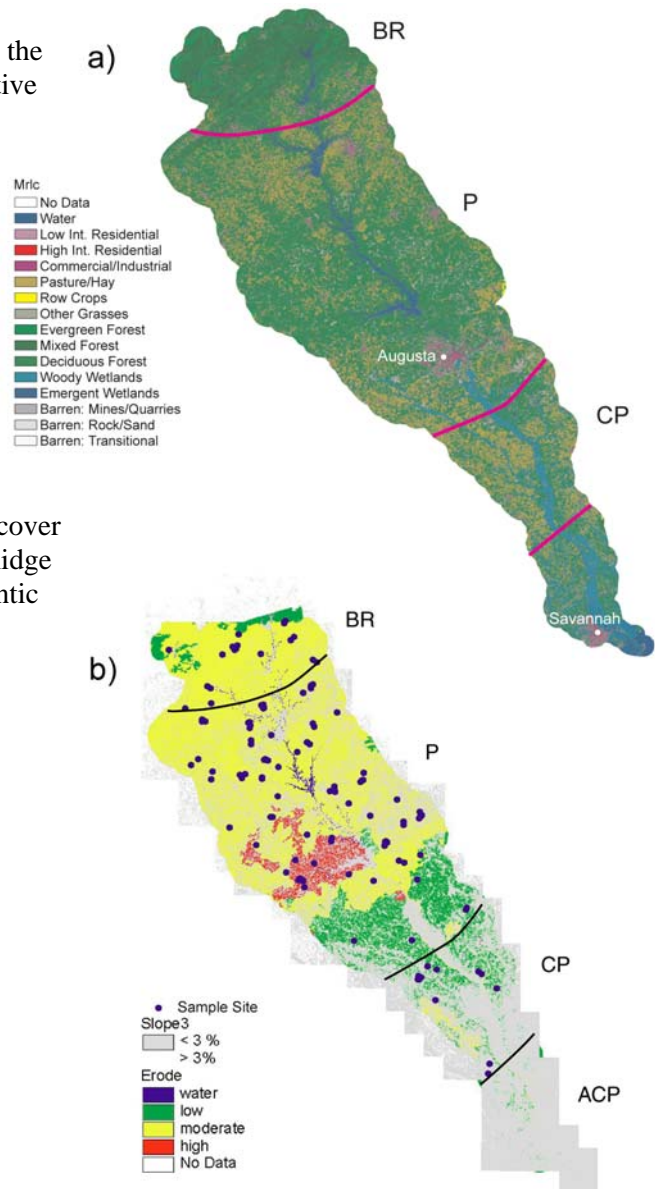


Figure 1. The Savannah River Basin ecoregion and MRLC classification (a) and the sample locations and erodibility classes (b).

Section 5

Statistical Analyses

Before presenting the canonical correlation analyses, detailed discussions about missing observations, collinearity, normality, and others are given in Appendix I. We used SAS™ for all the statistical analyses listed below.

5.1 Canonical Correlation Analysis

In canonical correlation analysis, the first attempt is to derive a linear combination (*canonical variate*) of the variables of each data set so that their correlation is maximized and is at least as large as the multiple correlation between any variable in one data set and all those from the other data set. For example, when surface water biota and landscape metrics were analyzed, the biological and landscape original variables is combined in a linear relationship, such as:

$$\text{Bio1} = \sum \varphi_i B_i$$

$$\text{Land1} = \sum \omega_i LS_i$$

where Bio1 and Land1 are the first pair of biological and landscape canonical variates and φ_i and ω_i are the coefficients (sometime referred to as weights) for each of the biological and landscape variables. B and LS are the biological and landscape original variables, respectively. Values of the coefficients, φ_i and ω_i , are determined so that the correlation between Bio1 and Land1 is maximized. The first pair of canonical variates always has the highest canonical correlation (r_k). The canonical correlation is considered as a multiple correlation of, for example, Bio1 with the landscape metrics, or, Land1 with the biota metrics. At this point, note that each variate (e.g., Bio1) can be considered a principal component (Gittens, 1985), and the covariance between the two variates (e.g., Bio1 and Land1) is maximum. If the canonical correlation is done on the standardized variables, each canonical variate is a principal component and the maximizing correlation and covariance are the same.

The analysis produces a second pair of variates (Bio2 and Land2) that is a linear combination of the original variables unlike the first pair; this pair of variates has the second highest canonical correlation. The analyses will continue to produce a number of pairs (k) until they are equal to the number of variables in the smaller data set. In other words, if the biological data set contains four variables and the landscape data set contains seven variables, then $k = 4$ pairs of variates will be produced. They are then sorted by their canonical correlation values, the first being the highest. In some references, k is referred to as the rank of the model. For simplicity and clarity, we will use the names below for each pair of canonical variates:

Bio1: first canonical biological variate,

Bio2: second canonical biological variate,

Chem1: first canonical chemical variate,

Chem2: second canonical chemical variate,

Land1: first canonical landscape variate, and

Land2: second canonical landscape variate.

In addition to canonical correlation, the analysis provides other statistics that can be used to fully understand and describe association. These are the canonical variate, square canonical correlation, canonical coefficient, and inter-set and intra-set structure correlations. Theory and development of canonical correlations are well stated in many references (Johnson and Wichern, 2002; Rencher, 1998; Hair et al., 1987; Gittins, 1985; Thorndike, 1978; Clark, 1975). A brief description of each statistic is given below.

5.1.1 Canonical Variate and Model Rank

The linear combination of the two data sets is termed the canonical variate. The number of pairs (k) of *canonical variate* is equal to the number of variables in the smaller data set. We examined the canonical correlation value and its significance level for each pair so the number of pairs to be used for interpretations in the final model can be selected. The number of significant pairs in a model is known as the rank of the model. Plotting the scores of variates in a pair describes visually the correlation between the linear combination of the two data sets (the variates).

5.1.2 Squared Canonical Correlation

The squared canonical correlation (r_k^2) is also known as eigenvalue or canonical root (Hair et al., 1987). The value of r_k^2 measures the proportion of variance of a canonical variate (e.g., Bio1) explained by the original variables of landscape metrics. The value of r_k^2 can be considered the multiple regression correlation between the two variates (e.g., Bio1 and Land1; Griffith and Amrhein, 1997), which measures the adequacy of the overall fitted model (Gittens, 1985). The summation of r_k^2 over k (the number of canonical pairs) is the *trace*, which measures the amount of variance that is shared and predicted by the two data sets. The final fitted model contains only the significant number (k) of canonical correlation. The quality of the model is measured by the percent trace (% $Trace = r_k^2 * 100 / \sum r_k^2$) and the percent cumulative trace (Table 2).

Table 2. Canonical correlation (r_k), canonical r-square (r_k^2), percentage trace (% $Trace = r_k^2 * 100 / \sum r_k^2$), and cumulative trace (%) of total (standardized) sample variance explained by each canonical variate and the probability of exceeding the critical value (F).

a) Biological and Landscape Metrics					
k	r_k	r_k^2	Trace %	Trace Cum. (%)	P>F
1	0.739	0.546	59.21	59.21	<0.0001
2	0.534	0.286	30.97	90.18	0.0008
3	0.267	0.071	7.71	97.89	0.2953
4	0.139	0.019	2.11	100.00	0.4653

b) Chemical and Landscape Metrics					
k	r_k	r_k^2	Trace %	Trace Cum. (%)	P>F
1	0.747	0.559	72.73	72.73	<0.0001
2	0.378	0.143	18.56	91.30	0.0134
3	0.259	0.067	8.70	100.00	0.0827

Table 2, Continued

c) Biological and Chemical Metrics					
k	r_k	r_k^2	Trace %	Trace Cum. (%)	P>F
1	0.543	0.353	59.62	59.62	<0.0001
2	0.480	0.231	39.01	98.62	0.0034
3	0.090	0.008	0.38	100.00	0.7451

5.1.3 Canonical Coefficients

Canonical coefficients are also known as canonical loadings or weights, and each canonical variate is a linear combination of the original variables weighted by their coefficients. The eigenvector associated with the (k^{th}) eigenvalue gives the weights for combining one set of variables (dependent). Weights for the other set of variables are produced by substituting these into the equation and solving. Hence, the loading measures the contribution of each of the original variables to the canonical variate (see coefficients in equations for the canonical variates on page 20). These weights were estimated so that the correlation between the two variates in a pair is maximized. Therefore, the values of these weights are expected to be different for different combinations. For interpretation of results, standard canonical coefficients (not raw canonical coefficients) were used for unification of units and scales of the original variables. To quantify the contribution of each variable to the canonical variates in a multivariate context, Rencher (1998) and Johnson and Wichern (2002) recommended using the standardized coefficients instead of the correlation between the original variable and canonical variate. For example, the coefficients ϕ_i 's ($\text{Bio1} = \sum \phi_i B_i$) reflect the joint contribution of the biota variables to the correlation between Bio1 and Land1.

It is important to note that when the sample size is small and the ratio of the number of variables to the number of observations is high, then the coefficients are unstable. More discussion about this issue is included in Appendix I.

5.1.4 Correlation Between the Original Variables and Canonical Variates

The correlation between the original variables and its canonical variates is known as the *intra-set structure correlation*. The intra-set correlation describes the amount of variance that is contributed by the original variable to their canonical variate. If the intra-set correlation is squared, it will give the amount of variance that is explained by its variate. Intra-set correlation indicates the strength of the association between the original variable and the canonical variates (which should give an indicator of the importance of the contribution of the original variable to the canonical variate). The square of the intra-set correlation would give the proportion of the variance of the canonical variate that is explained by the original variable.

The correlation between the original variables (e.g., AgPT) and the opposite (e.g., Land1) canonical variate is known as the *inter-set structure correlation*. This correlation is analogous to that of multiple correlation coefficients that describe the linear relationship between the original variable and the opposite canonical variate. For example, index of biological integrity (IBI) can be correlated to the water chemistry canonical variate. When the inter-set correlation value is squared, it will measure the amount (proportion) of variability in the original variable that is predicted by the opposite canonical variate. It is analogous to partial R^2 value. SASTM outputs the latter in the last part of SASTM canonical correlation analyses “squared multiple correlation in redundancy analysis” (see Table 3).

Table 3. Canonical correlation structure: correlation between the original variables and the canonical variate.

a) Biological and Landscape Metrics									
Variable	Bio1	Bio2	Land1	Land2	Variable	Land1	Land2	Bio1	Bio2
AgPT	-0.433	0.221	-0.320	0.118	Slope3	0.831	0.521	0.614	0.279
IBI	-0.324	0.267	-0.239	0.143	Soil_er	-0.151	0.884	-0.112	0.472
Hab	0.211	-0.816	0.156	-0.436	Pct_bar	-0.109	-0.556	-0.081	-0.297
Rich	0.431	-0.067	0.319	-0.036	Past_slp	-0.321	-0.669	-0.238	0.358
EPT	0.856	0.052	0.633	0.028					

b) Chemical and Landscape Metrics									
Variable	Chem1	Chem2	Land1	Land2	Variable	Land1	Land2	Chem1	Chem2
DO	0.891	0.078	0.666	0.029	Pct_bar	-0.524	0.211	-0.392	0.080
pH	-0.120	0.949	-0.090	0.358	Pct_for	0.073	0.871	0.054	0.329
EC	-0.844	0.037	-0.631	0.014	Past_slp	0.398	-0.249	0.297	-0.094
					Slope3	0.872	0.472	0.652	0.178

c) Chemical and Biological Metrics									
Variable	Bio1	Bio2	Chem1	Chem2	Variable	Chem1	Chem2	Bio1	Bio2
IBI	0.111	0.580	0.066	0.279	DO	0.714	-0.575	0.424	-0.276
Hab	0.254	0.053	0.151	0.025	pH	0.339	0.721	0.201	0.346
Rich	0.617	0.576	0.366	0.279	EC	-0.736	0.336	-0.437	0.162
EPT	0.901	0.065	0.535	0.031					

The correlation between the original variables and the canonical variates is more stable than the row or standardized canonical weights, i.e., coefficients (Gittens, 1985). Many have recommended, therefore, using correlations instead of this coefficient. Rencher (1998), however, verified the instability of a coefficient when a variable is added or deleted because it is a reflection of the mutual influence (multivariate) between variables in a canonical variate. The correlation between the original variable and its variate provide no information about the multivariate contribution of a variable to its variate. The relationship described by this correlation is a univariate relationship between the dependent and independent variables (Rencher, 1988, 1998, page 329). The square value of the correlation will measure the multiple correlation of that variable (e.g., IBI) with the other set of variables (e.g., the canonical variate of landscape variables). Therefore, correlations will not describe the joint contribution of the biota (dependent variables) to the canonical correlation with the landscape metrics (independent variables).

Correlations between the original variables and its variate and redundancy analyses are normally used to describe the contribution of the original variable to its or opposite canonical variate. With the above discussion, these statistics will not provide information in a multivariate relationship. Canonical correlations, square canonical correlations, and percent trace (Table 2) are the measures to be used to describe the multivariate association between the y's and x's.

5.1.5 Variables Entering the Canonical Correlation Analysis

Issues of missing values, collinearity, and others are described in detail in Appendix I, Supplementary Information. These issues may be of concern when applying canonical correlation analyses because of their possible effect on the number of observations and variables to be used in the analyses. In addition to the above issues, we eliminated any *nested structure* in a group of watersheds to remove a confounding effect. Sampling stations that were within the watershed support area of a downstream sampling station were omitted from analysis. The final data set for the canonical correlation consisted of:

- 1) Eighty-four sample sites with five biological (AgPT, IBI, Hab, Rich, and EPT) and four landscape (Past_slp, Pct_bar, Soil_er, and Slope3) metrics
- 2) Seventy-seven sample sites with three chemical (pH, DO, and EC) and four landscape (Pct_for, Pct_bar, Past_slp, and Slope3) metrics
- 3) Seventy-seven sample sites with three chemical (pH, DO, and EC) and four biological (IBI, Hab, Rich, and EPT) metrics

For variable descriptions, see Table 1.

5.1.6 Results

Pairwise Correlation of the Original Variables

- a) **Biological and Landscape Metrics:** The absolute correlation for the biological data was highest for the EPT and Rich (0.82), whereas, for landscape metrics it was 0.43 for Pct_bar and Past_slp. The highest correlation between the biological and landscape metrics was between EPT and Slope3 (0.54).
- b) **Chemical and Landscape Metrics:** The absolute value of correlation for the water chemistry data was highest for the DO and EC (0.51). A negative correlation was observed between DO and Pct_bar (-0.42), and between EC and Slope3 (-0.56). Positive correlation was found between pH and Pct_for (0.33). However, the highest correlation between the water chemistry and landscape metrics was between DO and Slope3 (0.58).
- c) **Chemical and Biological Metrics:** The highest absolute correlation between the water chemistry data and the computed biological parameters was between EC and EPT (-0.41) followed by DO and EPT (0.349).

Canonical Correlation

- a) **Biological and Landscape Metrics:** The canonical correlation values (r_k) are reported in Table 2a, in which the first two canonical variates were significant ($p \leq 0.0008$). The biological-landscape fitted model, therefore, consists of the first canonical (Bio1-Land1) and second (Bio2-Land2) variates. The linear relationship between the first two canonical variates can be visually examined in the scatter plot of the scores of the first ($r = 0.74$) and second ($r = 0.53$) canonical variates (Figure 2). Squared canonical correlations (r_k^2 ; Table 2a) express:
 - 1) The variation in the linear combination of the biological metrics that is attributable to the linear combination of the landscape metrics
 - 2) The proportion of the variance of Bio1 explained by the original variables of landscape metrics

3) The amount of shared variance between surface water biological and landscape original variables

The first and second canonical variates explain 0.55 and 0.29 of the variation, respectively, as compared to less than 0.09 for any of the remaining canonical variates. The sum of $r^2_k = 0.92$; Table 2a) measures the variance shared by all four pairs of canonical variates. The fitted model, which represents only the significant pairs (first and second), accounted for 90 percent (%Trace = $(0.55 + 0.29) * 100/0.92$; Table 2a) of that overall shared variance. Percent trace can be used as an index of the fitted model quality that measures prediction adequacy.

- b) Chemical and Landscape Metrics:** The first two canonical correlations are significant ($p \leq 0.0134$; Table 2b). The fitted model consists of the first two canonical variates only. The strength of the linear relationships for the first two pairs is 0.75 and 0.38, respectively. The percent variation of a canonical variate of water chemistry attributable to a landscape variate was 0.56 and 0.14 for the first and second pairs, respectively. Variance shared by all three pairs was 0.768 (the sum of the r^2_k in Table 2b). Our fitted model accounted for 91 percent (%Cum. Trace = $(0.56 + 0.14) * 100/0.768$) of the variance that is shared by all canonical variates. The quality index (91 percent) of the fitted model indicates that Chem1 and Chem2 are adequately predicted by surface water chemical parameters derived from landscape metrics (Land1 and Land2). The linear correlation between the chemical and landscape metrics for the canonical variates is shown in Figure 3.
- c) Biological and Chemical Metrics:** The first two canonical correlations are significant ($p \leq 0.0034$; Table 2c). Our fitted model consists of Bio1-Chem1 and Bio2-Chem2. The strength of correlation between Bio1-Chem1 is 0.59, and between Bio2-Chem2 is 0.48. The amount of variation in Bio1 attributable to Chem1 is 0.35, and the amount of variation in Bio2 attributable to Chem2 is 0.23. The variance shared by all three canonical variates is 0.59 (the sum of the r^2_k in Table 2c). Our fitted model accounts for 99 percent (%Cum. Trace = $(0.35 + 0.23) * 100/0.59$) of the variance that is shared by all the canonical variates. The quality index (99 percent) of the fitted model indicates the high quality of prediction. The linear correlation between the biological and chemical metrics for the canonical variates is shown in Figure 4.

Canonical Coefficient

For this section we list only the linear combination of the original variables for variates in the canonical pair(s) that is (are) significant ($\alpha < 0.05$). The canonical coefficient (or weight) indicates the relative contribution of each variable (standardized) to the canonical variate by magnitude and direction (positive or negative; see equations below). A coefficient value reflects the contribution of a variable in the presence of other variables that are in their canonical variate. This contribution has its effect on the opposite canonical variate as measured by the canonical correlation. Hence, multivariate context is more evident in coefficients than in correlation between the original variables and its canonical variate (more discussion in inter- and intra-set correlation).

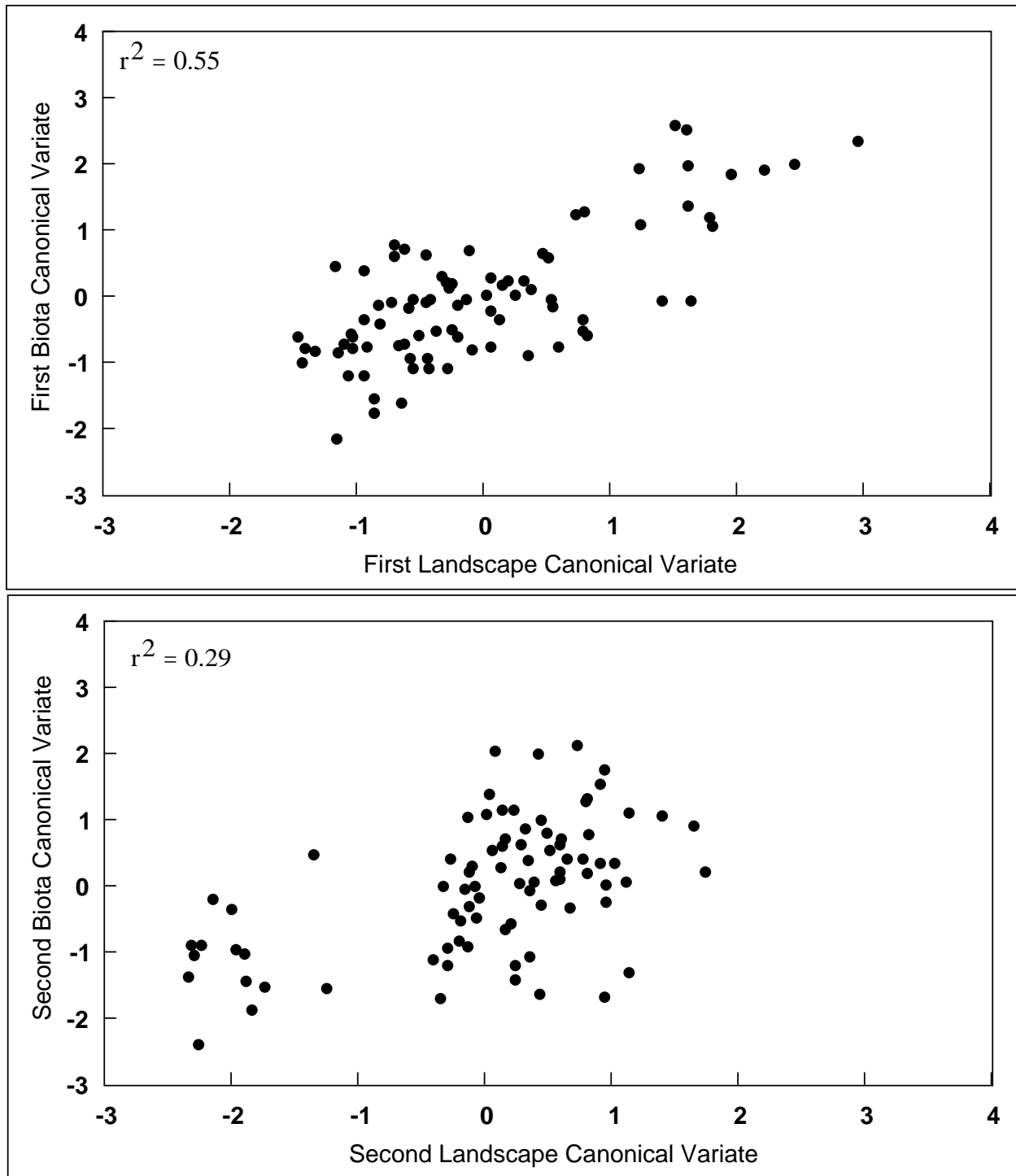


Figure 2. The scatter plot of the first two pairs of canonical variates for the landscape (Past_slp, Soil_er, Pct_bar, and Slope3) and surface water biological (AgPT, IBI, Hab, Rich, and EPT) data sets.

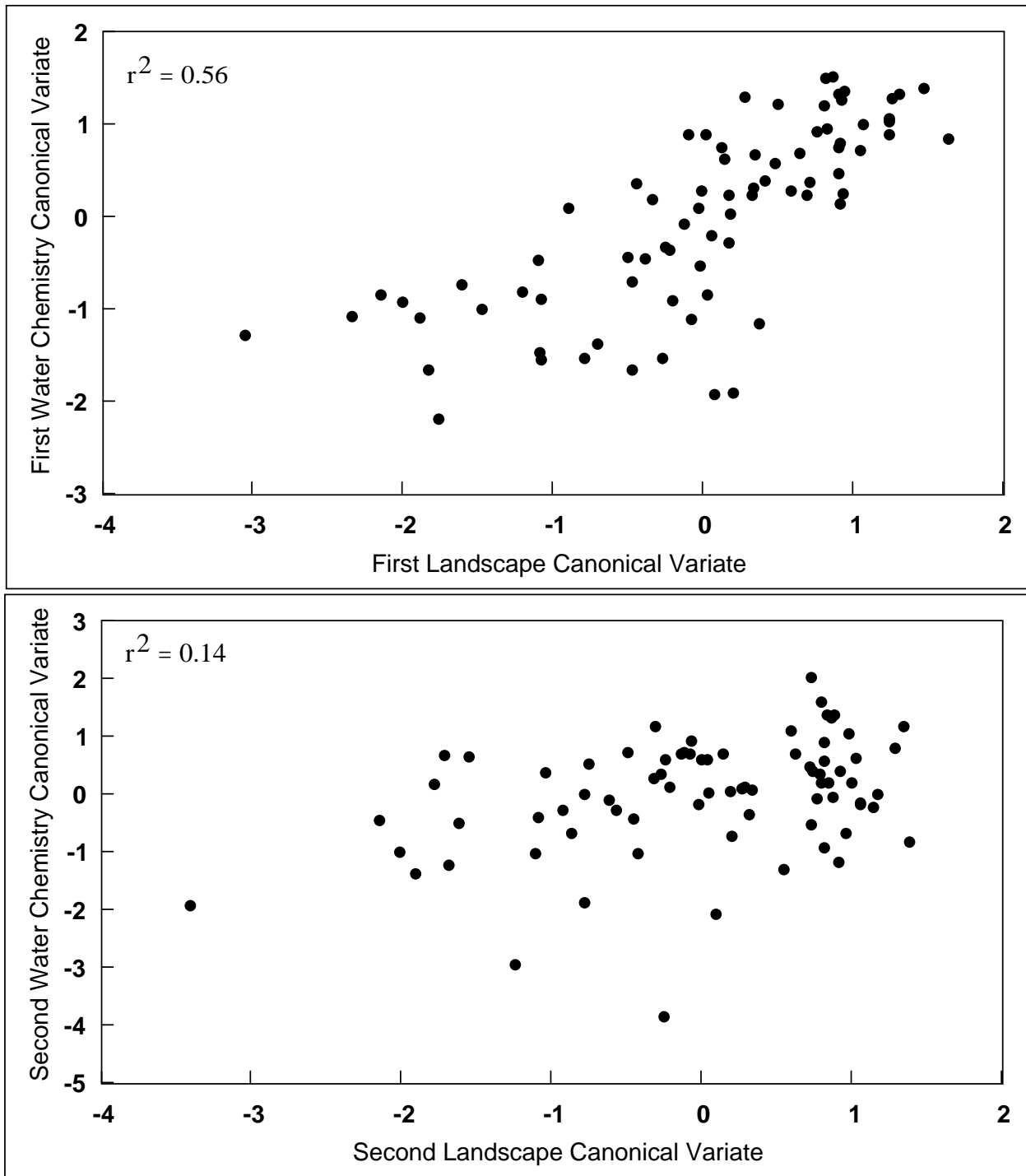


Figure 3. The scatter plots of the first two pairs of canonical variates for the landscape metrics (Past_slp, Pct_for, Pct_bar, and Slope3) and surface water chemistry (DO, pH, and EC) data sets.

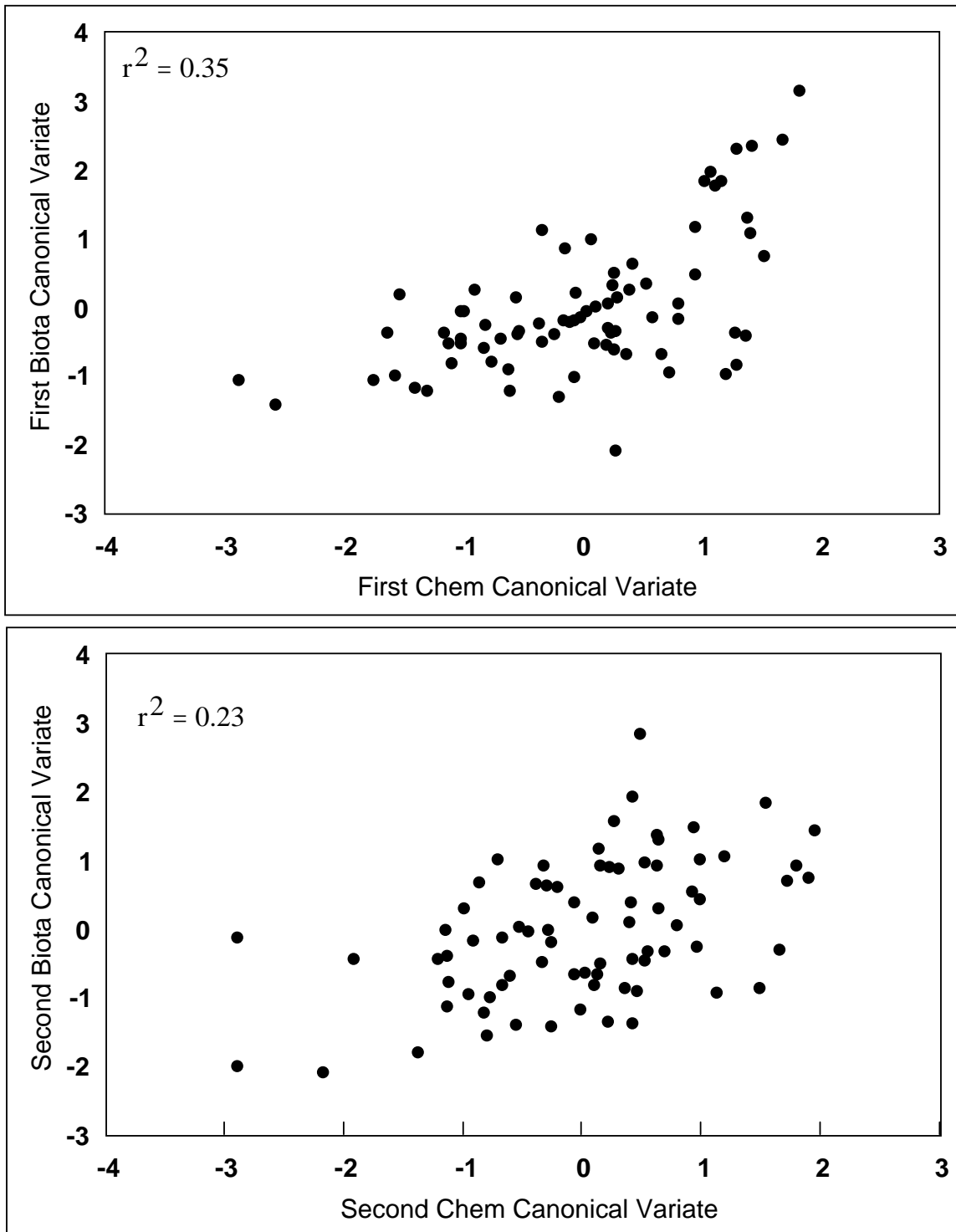


Figure 4. The scatter plots of the first two pairs of canonical variates for the surface water biota (IBI, Hab, Rich, and EPT) and surface water chemistry (DO, pH, and EC) data sets.

- a) **Biological and Landscape Metrics:** The contribution of the EPT and Hab are positive for the first biological canonical variate, and AgPT, IBI, and Rich are negative (see equations below). The contribution of the EPT is the highest (1.484). The landscape canonical variate is weighted negatively on Pct_bar, Past_slp, and Soil_er and positively on Slope3. The highest values in magnitude are for the Slope3 (0.954). The first canonical correlation reflects the relationships between the biota variate (EPT and Rich) and the landscape variate (Slope3 and Past_slp). When Slope3 increases, it causes an increase in EPT and a decrease in Rich; when Past_slp increases, AgPT increases. The second canonical correlation reflects the relationships between the Hab and Soil_er. When Soil_er increases, it causes a decrease in Hab. Land1 and Land2 represent Slope3 and Soil_er, respectively, whereas, Bio1 and Bio2 represent EPT/Rich and Hab, respectively.

$$\text{Bio1} = -0.207 * \text{AgPT} - 0.008 * \text{IBI} - 0.895 * \text{Rich} + 0.110 * \text{Hab} + 1.484 * \text{EPT}$$

$$\text{Land1} = -0.1604 * \text{Pct_bar} - 0.4315 * \text{Past_slp} - 0.337 * \text{Soil_er} + 0.954 * \text{Slope3}$$

$$\text{Bio2} = 0.229 * \text{AgPT} + 0.488 * \text{IBI} - 0.2347 * \text{Rich} - 0.942 * \text{Hab} + 0.651 * \text{EPT}$$

$$\text{Land2} = -0.0814 * \text{Pct_bar} + 0.368 * \text{Past_slp} + 0.654 * \text{Soil_er} + 0.251 * \text{Slope3}$$

- b) **Chemical and Landscape Metrics:** DO and EC are loaded with approximately the same magnitude on the first canonical chemical variate but with opposite signs (see the equations below), whereas, the second chemical variate is heavily weighted on pH. The coefficients for the first and second landscape variate are heavily weighted on Slope3 and Pct_for, respectively.

$$\text{Land1} = -0.341 * \text{Pct_bar} - 0.615 * \text{Pct_for} - 0.299 * \text{Past_slp} + 1.130 * \text{Slope3}$$

$$\text{Chem1} = 0.634 * \text{DO} - 0.032 * \text{pH} - 0.511 * \text{EC}$$

$$\text{Land2} = 0.415 * \text{Pct_bar} + 1.3702 * \text{Pct_for} + 0.743 * \text{Past_slp} - 0.204 * \text{Slope3}$$

$$\text{Chem2} = -0.232 * \text{DO} + 1.088 * \text{pH} - 0.399 * \text{EC}$$

- c) **Chemical and Biological Metrics:** Coefficients for Bio1-Bio2 and Chem1-Chem2 are given in the equations below. EPT and Rich are weighted heavily (positive) in Bio1 and Bio2 (1.434, 1.448). EC (-0.810) and pH (0.909) are the main contributors to Chem1 and Chem2, respectively.

$$\text{Bio1} = 0.396 * \text{IBI} - 0.564 * \text{Rich} + 0.047 * \text{Hab} + 1.434 * \text{EPT}$$

$$\text{Chem1} = 0.602 * \text{pH} + 0.281 * \text{DO} - 0.810 * \text{EC}$$

$$\text{Bio2} = 0.412 * \text{IBI} + 1.448 * \text{Rich} - 0.180 * \text{Hab} - 0.991 * \text{EPT}$$

$$\text{Chem2} = 0.909 * \text{pH} - 0.820 * \text{DO} - 0.377 * \text{EC}$$

In examining the signs of the coefficients in the equations above, signs are sometimes switched between the first and second equation for that canonical variate. Signs and magnitude of coefficients may change to maximize correlation between pairs in a canonical variate.

Canonical Structure (Canonical Redundancy Analyses)

Correlation of each original variable with its canonical variate or its opposite is defined as canonical structure (see below). These correlations do not reflect the multivariate relationships. That is, a correlation value describes a univariate relation between a variable and its canonical variate without considering the existence of other variables. The canonical structure is the most used and is cited frequently, but its usefulness in the multivariate context is among the lowest (Rencher, 1998; Johnson and Wichern, 2002).

Within Each Data Set (Intra-set Correlation):

- a) **Biological and Landscape Metrics:** The correlation between the original variables and their canonical variate is given in Table 3a. For the first canonical variate, the direction of contribution was positive for the Hab, Rich, and EPT, and it was negative for the IBI and AgPT (Table 3a). The highest correlation was with the EPT (0.856) followed by AgPT (-0.433) and Rich (0.431). Therefore, the biological canonical variate (Bio1) represents mainly EPT and Rich. The landscape original variables correlate negatively on Land1 except for Slope3 (0.83). Past_slp correlated second to the highest and related negatively to Land1. The canonical landscape variates (Land1 and Land2) explained 66 percent of the variation in landscape variables. The canonical biota variate (Bio1 and Bio2) explained 41 percent of the variation in biological variables.
- b) **Water Chemistry and Landscape Metrics:** Chem1 and Land1 both represent DO, EC, and Slope3 (Table 3b). Chem2 and Land2 were loaded mainly by pH and Pct_for (Table 3b). Chem1 and Land1 denote the common pattern between the two data sets that is contributed by Slope3 and DO. Chem2 and Land2 denote the common pattern between the two data sets that is contributed by Pct_for and pH. The canonical landscape variates (Land1 and Land2) account for 57 percent of the variation in landscape variables, and the canonical chemical variates (Chem1 and Chem2) account for 81 percent of the variation in water chemistry variables.
- c) **Water Chemistry and Biological Metrics:** The direction of correlation was positive for all biological variables with the first canonical biological variate (Bio1, Table 3c). The highest correlation was with EPT (0.90), followed by Rich (0.62). Therefore, the biological canonical variate represents mainly EPT and Rich. Bio2 represents mainly IBI and Rich. Rich is common in both Bio1 and Bio2 and was correlated positively with Bio2. The original chemical variables correlated positively, except for EC (-0.74) and DO (-0.58), with Chem1 and Chem2, respectively (Table 3c). The canonical biota variates (Bio1 and Bio2) account for 49 percent of the variation in biota variables, and the canonical chem variates (Chem1 and Chem2) account for 71 percent of the variation in water chemistry variables.

One important point to denote is the discrepancy in sign for some coefficients in a variate (page 20) to the correlation value (Table 3). The coefficient values for the IBI and Rich (page 20, Land-Bio) have opposite signs when compared to that of correlation (Table 3a). This is normally explained by the collinearity between variables. Pairwise correlation for variables included in canonical correlation should not exceed 0.9. For our data, the highest correlation was 0.82 (EPT and Rich) and all other correlations were less than or equal to 0.54.

Between Data Sets (Inter-set Correlation):

- a) **Biological and Landscape Metrics:** The direction and the strength of the correlations between the original biological variables and their canonical variates were different. EPT, Hab, and Rich were positively correlated with Land1 (0.633, 0.319, and 0.156), whereas, AgPT and IBI were negatively correlated with Land1 (-0.32, and -0.239; Table 3a). As pointed out earlier, Land1 was heavily

weighted by Slope3. Hence, EPT, Rich and Hab are related directly to the total watershed area with slopes greater than 3 percent. A higher value in Slope3 (percent total area regardless of landcover type) will enhance EPT and Rich in surface water; however, Past_slp negates them. The first two canonical landscape variates (Land1 and Land2) explain 18 percent of the variation in the first two canonical biota variates. The first two canonical biota variates (Bio1 and Bio2) explain 24 percent variation in the first two canonical landscape variates. The amount of variability explained by the opposite canonical variate is low, indicating that both landscape and biological variates are not good overall predictors of the other set of original variables.

- b) Water Chemistry and Landscape Metrics:** Chemical parameters responded differently to landscape metrics. The correlation between the original chemical variables and both Land1 and Land2 (Table 3b) indicated that there is a direct positive response of DO to the landscape, while pH and EC had the opposite response. The strongest relationship was for DO (0.67). The magnitude of the correlation between DO and Land1 is basically the response of DO to the topographical feature, Slope3. A high Pct_bar was associated with higher EC (Land1 and Chem1; Table 3b). All chemical variables were found to correlate positively with Land2, with the highest value being for the pH. The direction and magnitude of the correlation between Chem2 and Land2 indicate that a high water pH relates to a high percent forest (Pct_for) and a low percent total area in hay pasture on slopes greater than 3 percent (Past_slp). The canonical landscape variates (Land1 and Land2) explained 33 percent of the variation in water chemistry.
- c) Water Chemistry and Biological Metrics:** The relationship between water chemistry and biological parameters is of interest to this study. The strength of the relationship may reveal the adequacy of the chemical data as a surrogate for biological data, which would be cost-effective for future studies. The correlation between biota and chemical variates (Table 3c) indicates that there was a direct positive response of all biological data with Chem1 and Chem2. The strongest relationship was found for EPT and Rich (0.535 and 0.366) on Chem1, respectively. Water chemistry variables EC and DO correlated negatively with Bio1 and Bio2, respectively. The magnitude of the correlation of EC and DO with Bio1 and Bio2 (Table 3c) may indicate the lower abundance of the EPT and Rich. The canonical chemistry variates (Chem1 and Chem2) explained 15 percent variation of the biota.

R² and Canonical Redundancy Analysis

When the canonical correlation between the original variables and the corresponding variate (e.g., EPT and Land1; Table 3a) is squared, the results are known as the squared multiple correlation or R². For example, the square of 0.633 (correlation of EPT and Land1; Table 3) is 0.4002. That is, 40 percent of the variability in EPT alone was explained by the first landscape variate. The squared multiple correlation indicates that the first canonical variate of the biota has more predictive power for EPT (0.40), Rich (0.105), and AgPT (0.1025) than that for IBI (0.057) and Hab (0.024). Whereas, the second biota variate has more predictive power for the EPT (0.401) and Hab (0.215). The first canonical variate of landscape metrics is a fairly good predictor of Slope3 (0.377), poorer predictor for Soil_er (0.0125) and Past_slp (0.0564), and nearly useless for predicting Pct_bar (0.0065). The second canonical variate of landscape metrics is a fairly good predictor of Slope3 (0.455), Soil_er (0.2354), and Past_slp (0.1843) and is nearly useless for predicting Pct_bar (0.009).

For the chemistry and landscape data sets, the first and second landscape canonical variates are fairly good predictors of DO (R² = 0.44) and EC (R² = 0.40). For biota and chemistry data sets, the first and second chemical canonical variates are poor predictors of EPT (R² = 0.29) and Rich (R² = 0.21). Partial R² values can be found in the redundancy test SAS output.

Again, Rencher (1998) proved that R^2 values do not provide information in the multivariate context but rather in univariate.

5.1.7 Summary

Applying the canonical correlation demonstrated above resulted in a measure of the strength of the relationship expressed by the canonical correlation (r) between two sets of multiple variables, including stream biology and landscape variables. If the canonical correlation is significant, it indicates the existence of relationships between the two sets and these data sets are dependent. Coefficients or weights for each set of dependent (e.g., biota) and independent (landscape) variables in the canonical variate were determined to maximize the correlation between the linear combinations. By applying canonical analysis, it is possible to develop a number of independent canonical variates that maximize the correlation between the linear composites of sets of biota and landscape metrics.

Statistics from SASTM outputs can easily be used to describe and understand the relationships. Many references cautioned using some of these statistics, such as the coefficients or weights (Rencher, 1998; Hair et al., 1987; Gittens, 1985), because they are sensitive to the variability between observations. Others recommended using the inter-set and intra-set correlations to describe the variance shared by the two data sets. We made the effort to use and describe all the statistics to show the pattern of relationships between the landscape and water chemistry and biological properties.

Within the traditional framework in fitting any regression model, a high R^2 value is a good indicator of a successful and well-received research paper. Many models may fit the data well (Hair et al., 1987) but have a low R^2 value. Other statistics, as an alternative, may give a better representation. Field data, especially environmental data, rarely have a high R^2 value, yet the fitted model may describe a valid physical phenomenon and a pattern of relationships to other variables. When an attempt was made to describe the amount of variability of biota that was explained by the landscape metrics, the R^2 value was not high. However, the canonical correlation analyses for these data sets revealed a useful pattern of relationships between landscape metrics and surface water quality. High correlations between the variates, the quality index of the fitted model, the standardized coefficient of the original variables in the canonical variate, and the correlation between the original variables and their own or opposite canonical variates helped to reveal an interesting pattern of relationships. Canonical correlation was used here for exploring relationships and not for predictive purposes.

The landscape-biological model indicated three major contributing variables: the percent of total area with slopes greater than 3 percent (Slope3), EPT (an indicator of three microinvertebrate genera), and Rich (an index of microinvertebrate species richness). Within this model, the landscape variable, pasture on slopes greater than 3 percent, was the second highest landscape contributor, with a negative relationship to the biological variables. Slopes greater than 3 percent were also a major contributing landscape metric in the landscape-water chemistry model; the major water chemistry contributing variable in this model was dissolved oxygen (DO). In the chemistry-biota model, EPT and Rich were again the major contributing biological variables, while EC and pH were the major contributing water chemistry variables, with EC negatively related to biological variables.

The strength of landscape and chemistry relationships was higher than those for landscape and biota. In circumstances where biota data were unavailable or could not be measured because of costs or other restrictions, the condition of the stream biology can be assessed using chemistry data as surrogates with a moderate degree of confidence.

In conclusion, canonical correlation analyses indicated that increased slope (indicating complex topography, generally occurring in the mountainous areas of the Savannah River Basin) is associated with

increased macroinvertebrate species richness and higher DO concentrations, while the percentage of landcover in pasture on slopes greater than 3 percent is associated with increased conductivity and a decline in aquatic species richness.

5.2 Partial Least Square (PLS) Analysis

This method is widely used in chemometrics to describe how structural variations in chemical compounds affect biological activity. It involves the prediction of chemical form from spectroscopy data, where several hundred wavelengths and a smaller number of chemical samples (Owen, 1988) are the norm. In quantification of molecular modeling, a large number of independent variables (>1,000) are normally obtained with respect to the number of samples (10-100).

In ecology, we are interested in describing how the structural variation in landscape metrics may affect surface water biological and chemical properties. PLS can be of use in landscape studies utilizing the variation in both independent and dependent data sets. With the advances in GIS technology, obtaining landscape metrics for the past and present in a study area is no obstacle. This may or may not be, especially for the past, the case in obtaining the same amount of surface water data. Past data for surface water quality may be incomplete (missing/nonmeasured) or have a small sample size. Therefore, this does not preclude the use of PLS.

PLS is a multivariate analysis that decomposes the original data into a set of orthogonal factors. PLS is specifically close to canonical correlation. Also, it is a generalization of multiple regression. In contrast to other multivariate analyses (e.g., canonical correlation analysis), PLS is a linear predictive model for the dependent and independent data sets. In multiple regression, the variation of values for the independent variables is used to predict the dependent variable. PLS, on the other hand, uses the variation in both the independent and dependent data sets to predict the dependent variables. In multiple regression, as well as in canonical correlation, a strong collinearity in independent variables and a large number of independent variables compared to that of observations are potential problems and have to be dealt with before running the analyses. In addition, multivariate cannot handle missing data. That is, an observation will not be included in the analyses when a value is missing for any variable, either dependent or independent. PLS, however, can analyze and predict for data sets with missing values (as discussed below), collinear independent variables, and with independent variables that have no structure in the behavior (“noisy”).

PLS is used to analyze two data sets (e.g., the previously described biological and landscape data sets). Both data sets are first centered and scaled (e.g., Bio0 and Land0), then a linear combination is composed of the dependent variables ($v = \text{Bio0} * w$; v is the score and w is the weight) and the independent variables ($u = \text{Land0} * t$; u is the score and t is the weight). The linear composition of each data set is then built to maximize the covariance between them. A second linear composition is built using the residuals from the first factor to find the linear combinations of both data sets to maximize their covariance. The process is repeated by taking residuals from previous factors and producing $n-1$ factors. For example, if the number of sites (observations) is 89, then 88 factors will be produced.

Values of Euclidean distance from each point to the model prediction both for dependent and independent variables can be plotted. Plots of distances from the two data sets (x 's and y 's) to the model prediction can help with visualization of an outlier, consequently identifying the observation(s) that is(are) inconsistent with general behavior of the data.

5.2.1 Fitting the Model

PLS produces a number of significant factors using a cross validation (CV) method. Not all $n-1$ factors contribute significantly to prediction, hence, cross validation is used to select the first few factors that are significant. Cross validation starts by dividing the data into groups (five to nine groups (Wold, 1995)). If, for example, the data were divided into five groups, one group (test data) is left out and the model is fitted to the other four groups (training data). In SASTM there are five CV grouping methods: one, split, block, random, and test set. One, also known as “leave one out,” fits the model on $n-1$ observations and uses the one left out for validation; however, it is not recommended (Wold, 1995). An important point to consider is PLS does not require large training and test data sets. The fitted models can be tested using the test data sets, and the predicted values are compared to that observed using predictive residual sum of square (PRESS) to assess the predictive ability of the model. The model with the minimum root mean PRESS is the one to use (Table 4). However, a model with less number of factors that have a root mean PRESS value may not be significantly different than a model with the absolute minimum root PRESS. Van der Voet (1994) suggested using statistics (analogous to Mahalanobis) to test differences between root mean PRESS values of models. The chosen model will be the one with fewer factors with a PRESS value that is not significantly different than that of minimum PRESS.

In SASTM, cross validation is done by using options CVTEST (see SASTM for PLS statements in Appendix III). SASTM gives the root mean PRESS for each model and the significance level of the test to indicate whether that model is different from the one with the absolute minimum PRESS.

5.2.2 Diagnostic Checking and Variable Influence on Projection (VIP)

We plotted each factor's scores and weights to examine the strength of relationships and irregularities. *Scores* of both data sets ($X_$ and Y_Scores) show irregularity in grouping and outliers in observations/sites (Figures 5 and 6), whereas, *weights* show irregularity in grouping and outliers for the independent variables. Weights are used in determining the variable influence on projection (VIP) that can be used to select the most important variables. VIP is a statistic that PLS produces showing the contribution of the independent variable to the model (Table 5). Sometimes in regression, when the absolute value of the coefficient is small, the contribution of that independent variable to the predicted value is considered small. Consequently, that variable is deleted from the model. This may not be the case in PLS using VIP. An independent variable may have a small value of a coefficient but may have a large VIP, implying that this independent variable is important and contributes significantly to the prediction and, therefore, has to be kept in the model. If the values of the VIP and the coefficient both are small, that variable may be deleted from the model. Deletion of a variable from a model should also be based on the biological importance or other scientific judgments (Nash and Bradford, 2001). Wold (1995) indicated a VIP value of less than 0.8 is considered to be small.

Table 4. The significant factors for the preliminary PLS model for the surface water biological (5) and landscape metrics (26).

Split-sample Validation for the Number of Extracted Factors			
No. of Extracted Factors	Root Mean PRESS	t²	P > t²
0	1.0616	10.7153	0.0440
1	0.9962	0	1.0000
2	1.0925	10.6000	0.0280
3	1.1459	8.0576	0.0780
4	1.1561	11.0342	0.0280
5	1.1535	12.3791	0.0130
6	1.2262	9.6425	0.0370
7	1.2402	9.0255	0.0520
8	1.2315	8.7078	0.0690
9	1.2455	6.7807	0.2090
10	1.2860	6.1018	0.2840
11	1.2498	7.0152	0.1840
12	1.2688	6.4794	0.2350
13	1.2766	6.8627	0.1970
14	1.3077	7.5437	0.1340
15	1.3759	7.5264	0.1240

Minimum root mean PRESS	0.9962
Minimizing number of factors	1
Smallest number of factors with P>0.1	1

Percent Variation Accounted for by PLS Factors				
No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	29.6554	29.6554	17.6607	17.6607

Table 5. PLS regression coefficients and VIP values for the preliminary model (5 biota and 26 landscape metrics).

Predictor	AgPT	EPT	IBI	Hab	Rich	VIP
Ag_hi	-0.0039	0.0047	-0.0009	0.0028	0.0034	0.1092
Ag_slp_hi	-0.0048	0.0057	-0.0011	0.0034	0.0041	0.1336
Ag_mod	0.0536	-0.0646	0.0121	-0.0385	-0.0466	1.5032
Ag_slp	0.0438	-0.0529	0.0099	-0.0315	-0.0381	1.2303
Ag_slp_mod	0.0448	-0.0540	0.0101	-0.0322	-0.0390	1.2571
Bar_slp_hi	-0.0062	0.0075	0.0014	0.0044	-0.0054	0.1734
Bar_slp_mod	-0.0087	0.0104	-0.0020	0.0062	0.0075	0.2429
Crop_slp	0.0472	-0.0569	0.0107	-0.0340	-0.0411	1.3253
Crop_slp_mod	0.0504	-0.0608	0.0114	-0.0363	-0.0439	1.4150
Past_slp	0.0378	-0.0455	0.0085	-0.0272	-0.0328	1.0599
Pct_bar	-0.0095	0.0115	-0.0022	0.0068	-0.0083	0.2668
Pct_crop	0.0321	-0.0387	0.0073	-0.0231	-0.0279	0.9013
Pct_for	-0.0539	0.0650	-0.0122	0.0388	0.0469	1.5131
Pct_past	0.0465	-0.0561	0.0105	-0.0334	-0.0404	1.3046
Pct_urb	0.0219	-0.0264	-0.0049	-0.0157	-0.0190	0.6133
Pct_wet	-0.0075	0.0091	0.0017	0.0054	0.0066	0.2117
Pct_wtr	0.0254	-0.0306	0.0057	-0.0183	-0.0221	0.7131
PwrPipTI	-0.0054	-0.0066	0.0012	-0.0039	-0.0047	0.1527
Slope3	-0.0408	0.0491	-0.0092	0.0293	0.0354	1.1437
Slp_mod	-0.0047	0.0056	-0.0011	0.0034	0.0041	0.1306
Sd_slp	-0.0624	0.0752	-0.0141	0.0449	0.0543	1.7512
Mean_slp	-0.0617	0.0744	-0.0140	0.0444	0.0536	1.7309
Soil_er	0.0341	-0.0411	0.0077	-0.0245	-0.0297	0.9571
Strmden	-0.0365	0.0440	-0.0083	0.0262	0.0317	1.0241
TotRoad30	0.0082	-0.0099	0.0019	-0.0059	-0.0072	0.2307
TotRoadWS	0.0229	-0.0276	0.0052	-0.0165	-0.0199	0.6418

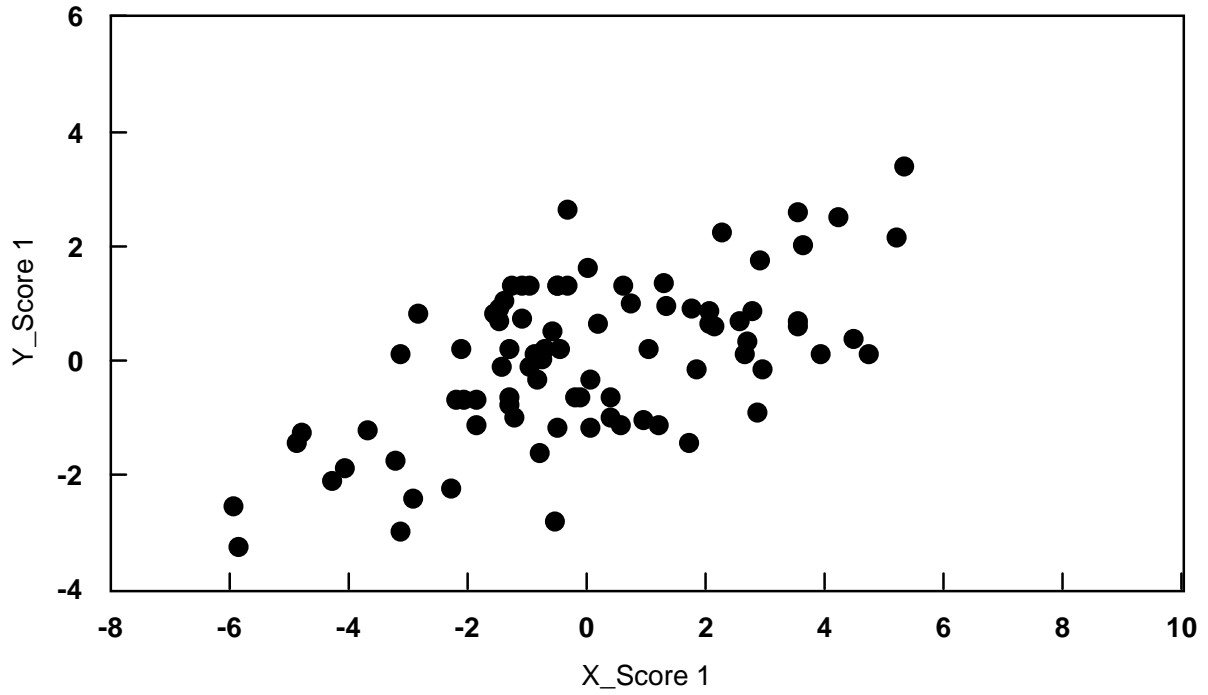


Figure 5. Plot of X_Scores (u) and Y_Scores (v) for factor 1 for each observation from the PLS model of 5 biota and 26 landscape metrics.

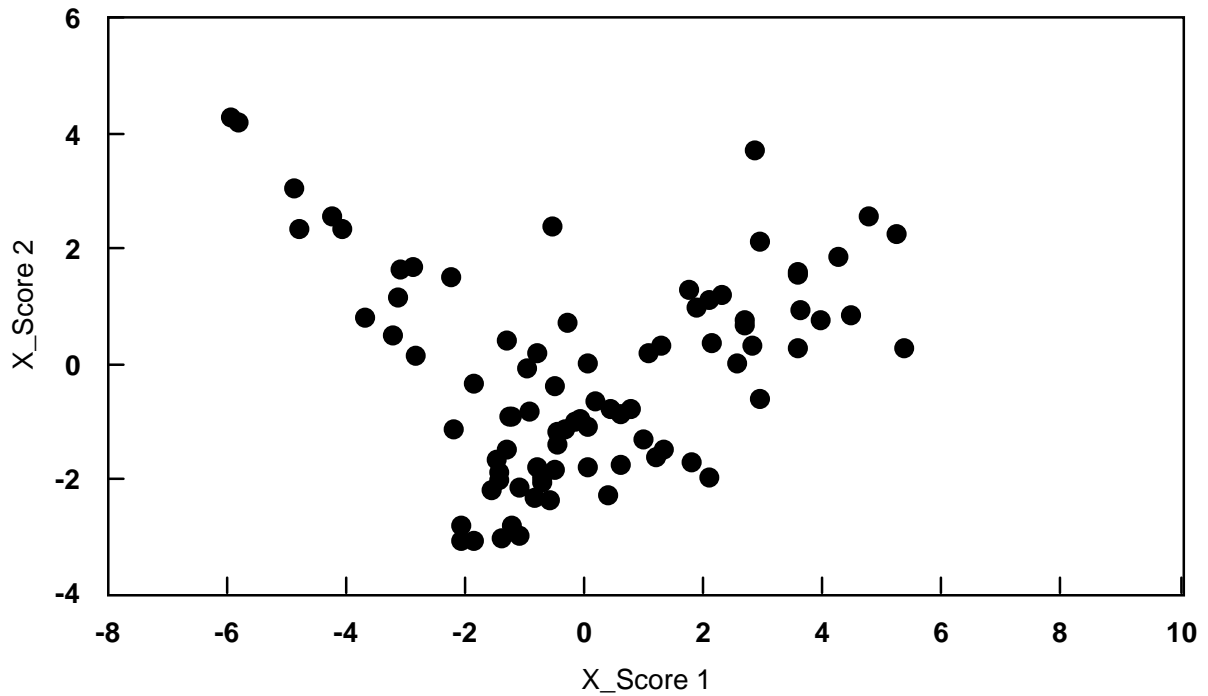


Figure 6. X_Scores from the first and second PLS factors for each observation.

5.2.3 PLS Analyses

5.2.3.1 Savannah River Basin

PLS is more applicable than canonical correlation for the Savannah River data sets because of the way that it deals with the number of variables compared to the number of observations, missing observations of some dependent variables, outliers, and collinearity. For our analyses, therefore, PLS was used for different data sets to fully capture its application to ecological data. Our analysis procedure consisted of the following three steps:

1. We used 26 landscape and 5 biological metrics (Table 1), including the missing values. The model was reduced to include independent variables (landscape metrics) with $VIP > 0.8$ and a diagnostic check of the model was completed for finalization.
2. The final model from (1) above was analyzed by ecoregion.
3. For the “Piedmont” ecoregion, missing values were excluded, and the best model was found. This model was used to predict for sites that have landscape metrics but no measured biological data.

All Data

The five dependent variables (AgPT, EPT, Hab, Rich, and IBI) and the 26 landscape metrics were analyzed. The cross validation was performed by dividing the data ($n = 86$) into two groups: training and testing. The testing data consisted of extracting every 9th value (*Split(9)* option in SASTM). Hence, the training data consisted of 78 observations, and the test data consisted of 8 observations. Using training and test data helps to identify the number of significant factors to include in the final model by cross validation. The model comparison test, which compares each model to that with the absolute minimum PRESS, is based on re-randomization of the data (CVTEST option in SASTM). The number of significant factors and the percent variation explained by the dependent and independent variables are given in Table 4. Estimates of the PLS regression coefficient and VIP values (Table 5) were examined. The PLS factors accounted for 18 and 30 percent of the variation for the biota and landscape data sets, respectively. Y_Score and X_Score values for the first and second factor show the strength of relationships and the distribution of sites (Figure 5), and there is no particular clustering pattern in these sites (Figure 6).

The plot of X-Weights for the first and second factors indicates which landscape metrics (predictors) are most represented in each factor (Figure 7). It's clear that, for example, the percent of erodible soil (Soil_er), and the percent of total area with slopes >3 percent (Slope3) and stream density (Strmden) have high (absolute) weights for both factors, while percent forest (Pct_for) is most represented in factor 1. Those variables that cluster near the origin have low weights on both factors and do not contribute much to the predictive capability of the model. Those variables that cluster near each other indicate equal weight on a factor. The VIP represents the importance of the variable in fitting the PLS for both the dependent and independent variables. The contribution of some of these 26 landscape metrics are low based on the VIP (Ag_hi, Ag_mod, Bar_slp_hi, Bar_slp_mod, Pct_bar, Pct_urb, Pct_wet, Pct_wtr, Slp_mod, Totroad30, and TotroadWS; Table 5). Landscape metrics such as Ag_mod, Ag_slp, and Ag_slp_mod have similar VIP values, hence, one may choose one landscape metric to describe agriculture on areas with slopes >3 percent or with moderate erodible soil. The same can be done for the variables Crop_slp, Crop_slp_mod, and Past_slp.

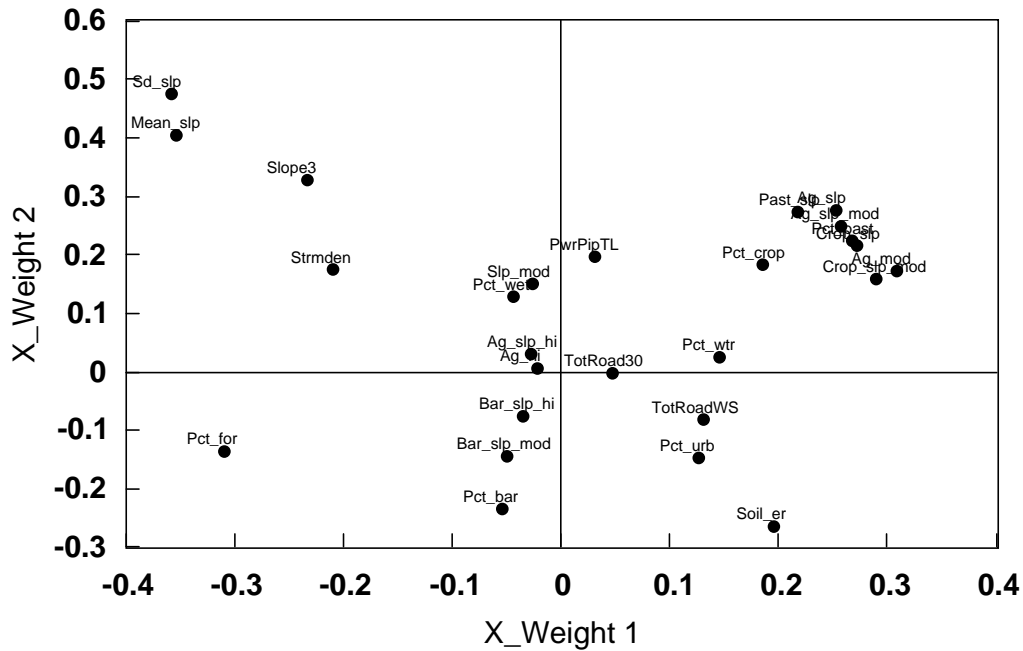


Figure 7. Plot of first and second X_Weights for the PLS model of the 5 biota and 26 landscape metrics.

The model may perform better if we include only landscape variables with $VIP > 0.8$ (Table 5). Hence, we ran another model that contained only 14 landscape metrics. The final model had two significant factors, which explained 22 and 72 percent of the variation in the biological (dependent) and landscape metrics (independent), respectively (Table 6). The importance of the 14 landscape metrics was all high ($VIP \geq 0.8$; Table 7), except for Pct_crop, Soil_er, and Strmden ($0.70 \leq VIP < 0.8$). X_ and Y_Scores and weights for the final model are shown in Figures 8 and 9. From Table 7 and Figure 8, a combination of percent agriculture, crop, and pasture were clustered in a group. This group of landscape metrics in addition to slopes, percent of total area with slopes > 3 percent (Slope3), stream density (Strmden), percent forest (Pct_for), and erodible soil (Soil_er) is contributing the most to explaining the variability in the biota in surface water. The contribution of one or many landscape metrics to the variability in the surface water biota may be different between ecoregions, which is investigated below in detail.

AgPT increased as the agriculture on areas with slopes > 3 percent and erodible soil increased but AgPT decreased as the percent forest, slopes, and stream density increased. AgPT also decreased as Pct_crop increased, further indicating the importance of slope and soil erodibility in relationships to surface water quality.

In the final model, outliers were examined by plotting the x-distance and the y-distance for each site (Appendix I, Figures I-2a and b). There was no evidence of outliers in the data.

Table 6. The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the final model.

Blocked Cross Validation for the Number of Extracted Factors			
No. of Extracted Factors	Root Mean PRESS	t²	P > t²
0	1.0616	14.7959	0.0040
1	0.9819	11.4172	0.0260
2	0.9686	6.6464	0.2590
3	0.9561	0	1.0000
4	0.9780	8.8522	0.0710
5	0.9769	7.5310	0.1780
6	0.9828	10.2828	0.0550
7	0.9813	6.3798	0.2890
8	0.9845	6.0714	0.3310
9	0.9857	5.1085	0.4540
10	0.9987	5.9617	0.3420
11	1.0221	8.9084	0.1120
12	1.0778	7.6985	0.1590

Minimum root mean PRESS	0.9561
Minimizing number of factors	3
Smallest number of factors with P>0.1	2

Percent Variation Accounted for by PLS Factors				
No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	51.4485	51.4485	17.4151	17.4151
2	20.9930	72.4414	5.0071	22.4223

Table 7. Regression coefficient values for the 5 biota and VIP for landscape metrics in the final PLS model.

Predictor	AgPT	EPT	IBI	Hab	Rich	VIP
Ag_mod	0.0597	-0.0437	0.0134	-0.0345	-0.0348	1.0473
Ag_slp	0.0524	0.0008	0.0063	-0.0226	-0.0106	0.9551
Ag_slp_mod	0.0526	-0.0083	0.0076	-0.0245	-0.0153	0.9452
Crop_slp	0.0543	-0.0199	0.0094	-0.0275	-0.0216	0.9632
Crop_slp_mod	0.0560	-0.0403	0.0125	-0.0322	-0.0323	0.9826
Past_slp	0.0461	0.0095	0.0043	-0.0182	-0.0049	0.8548
Pct_for	-0.0586	0.0596	-0.0155	0.0371	0.0427	1.0291
Pct_crop	-0.0384	-0.0041	0.0053	-0.0175	-0.0102	0.6921
Pct_past	0.0539	-0.0186	0.0092	-0.0270	-0.0209	0.9566
Slope3	-0.0342	0.1601	-0.0266	0.0459	0.0887	0.9759
Soil_er	0.0283	-0.1246	0.0209	-0.0365	-0.0694	0.7718
Strmden	-0.0337	0.1105	-0.0196	0.0361	0.0634	0.7626
Sd_slp	-0.0522	0.2319	-0.0389	0.0677	0.1291	1.4333
Mean_slp	-0.0534	0.2120	-0.0362	0.0643	0.1192	1.3552

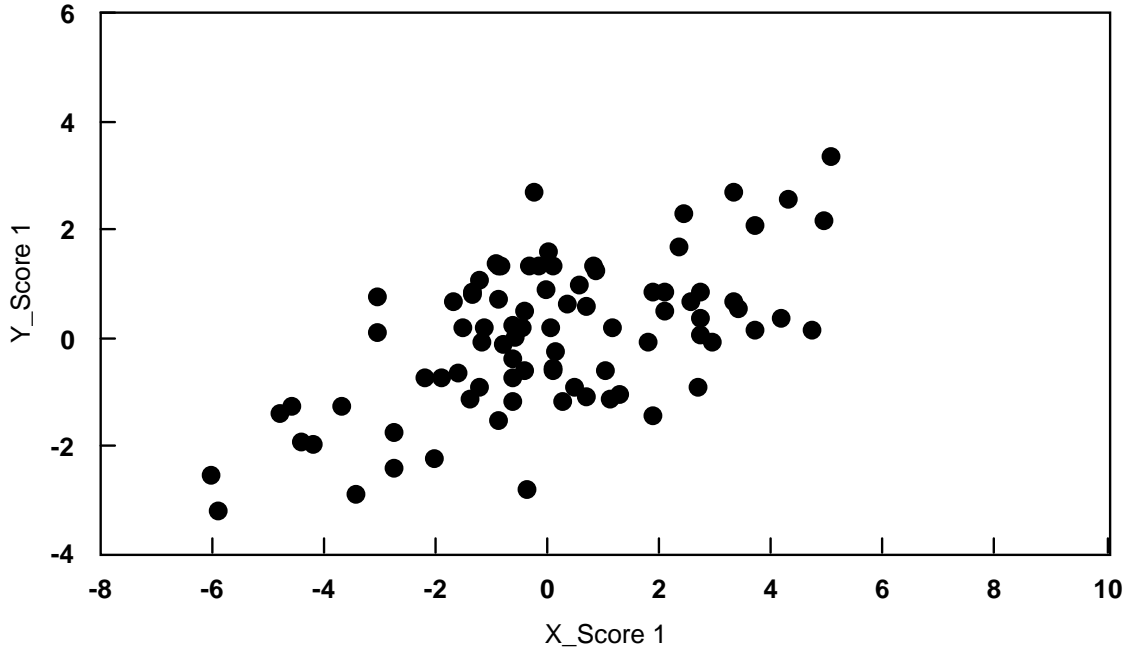


Figure 8. Plot of X_ and Y_Scores for the first factor of the final PLS model of the 5 biota and 14 landscape metrics.

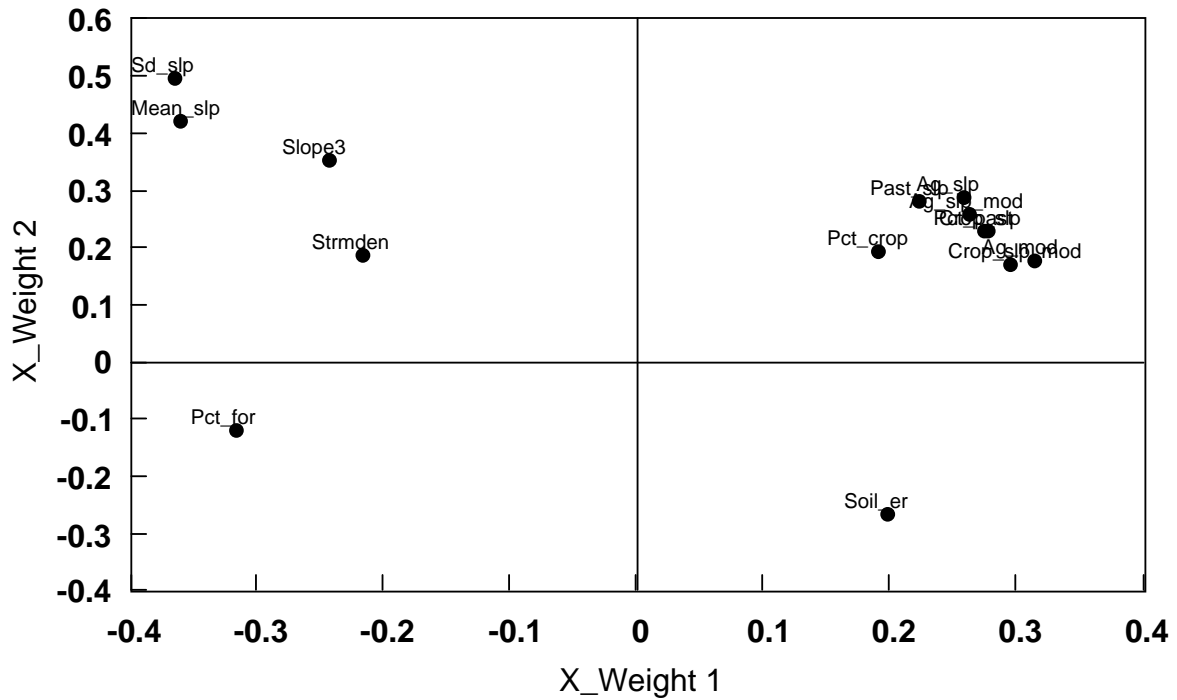


Figure 9. Plot of first and second X_Weights for the first PLS final model of the 5 biota and 14 landscape metrics. Variables included in the cluster in the upper right quadrant are Ag_mod, Ag_slp, Ag_slp_mod, Crop_slp, Crop_slp_mod, Past_slp, and Pct_past.

By Ecoregion

The final model with the 14 independent variables was performed by ecoregion.

- a) **Blue Ridge (BR):** There was only one significant factor that accounted for 21 and 74 percent of the variability for the biota and landscape metrics, respectively (Table 8). Ag_mod, Ag_slp, Ag_slp_mod, Past_slp, percent forest (Pct_for), percent pasture (Pct_past), total area with slopes >3 percent (Slope3), Sd_slp, and Mean_slp (VIP >1) were the most important variables followed by the percent crop (0.8 <VIP <1.0). Stream density, erodible soil (Soil_er), crops on slopes >3 percent (Crop_slp), and crops on areas with slopes >3 percent with moderately erodible soil (Crop_slp_mod) were less important (VIP <0.8; Table 9). A diagnostic check for the final model indicated a reasonable fit (scores of the first factor, Figure 10). The BR ecoregion is characterized by mountainous terrain, predominantly covered in evergreen forest. Barren areas are mainly of two types: transitional areas where the natural forest cover has been removed and mines. Soils are low-to-moderately erodible. Only a small percentage of the total landcover is in agriculture, predominantly pasture, and there are several small urban areas. In total, anthropogenic landcover types account for less than 10 percent of the land area. The results above indicated that percent forest, forms of agriculture, and topographic features are the driving elements affecting surface water quality.

Table 8. The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the Blue Ridge (BR) ecoregion (n = 19).

Blocked Cross Validation for the Number of Extracted Factors			
No. of Extracted Factors	Root Mean PRESS	t ²	P > t ²
0	1.3620	9.5030	0.0370
1	1.2393	0	1.0000
2	1.3086	5.7027	0.3610
3	1.5059	4.4335	0.5800
4	1.5583	5.3579	0.4160
5	1.7651	8.3547	0.0380
6	1.9155	7.8350	0.0830
7	2.0702	8.1628	0.0650
8	3.2720	7.2545	0.1120
9	3.8187	7.5213	0.0940
10	6.8551	6.8121	0.1920
11	119.2322	7.2010	0.0950
12	119.2322	7.2010	0.0950

Minimum root mean PRESS		1.2393
Minimizing number of factors		1
Smallest number of factors with P>0.1		1

Percent Variation Accounted for by PLS Factors				
No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	74.1617	74.1617	21.0517	21.0517

Table 9. Regression coefficient and VIP values for biological and landscape metrics values for the Blue Ridge (BR) PLS model.

Predictor	AgPT	EPT	IBI	Hab	Rich	VIP
Ag_mod	0.0434	-0.0562	0.0213	-0.0216	-0.0445	1.0343
Ag_slp	0.0450	-0.0582	0.0221	-0.0223	-0.0462	1.0723
Ag_slp_mod	0.0450	-0.0582	0.0221	-0.0224	-0.0462	1.0725
Crop_slp	0.0309	-0.0400	0.0152	-0.0154	-0.0317	0.7371
Crop_slp_mod	0.0310	-0.0401	0.0152	-0.0154	-0.0318	0.7384
Past_slp	0.0465	-0.0602	0.0229	-0.0231	-0.0477	1.1088
Pct_for	-0.0476	0.0616	-0.0234	0.0236	0.0488	1.1340
Pct_crop	0.0362	-0.0469	0.0178	-0.0180	-0.0372	0.8631
Pct_past	0.0431	-0.0557	0.0212	-0.0214	-0.0442	1.0267
Slope3	-0.0468	0.0605	-0.0230	0.0232	0.0480	1.1152
Soil_er	0.0182	-0.0235	0.0089	-0.0090	-0.0186	0.4329
Strmden	-0.0307	0.0397	-0.0151	0.0152	0.0315	0.7312
Sd_slp	-0.0554	0.0717	-0.0272	0.0275	0.0568	1.3201
Mean_slp	-0.0517	0.0669	-0.0254	0.0257	0.0531	1.2326

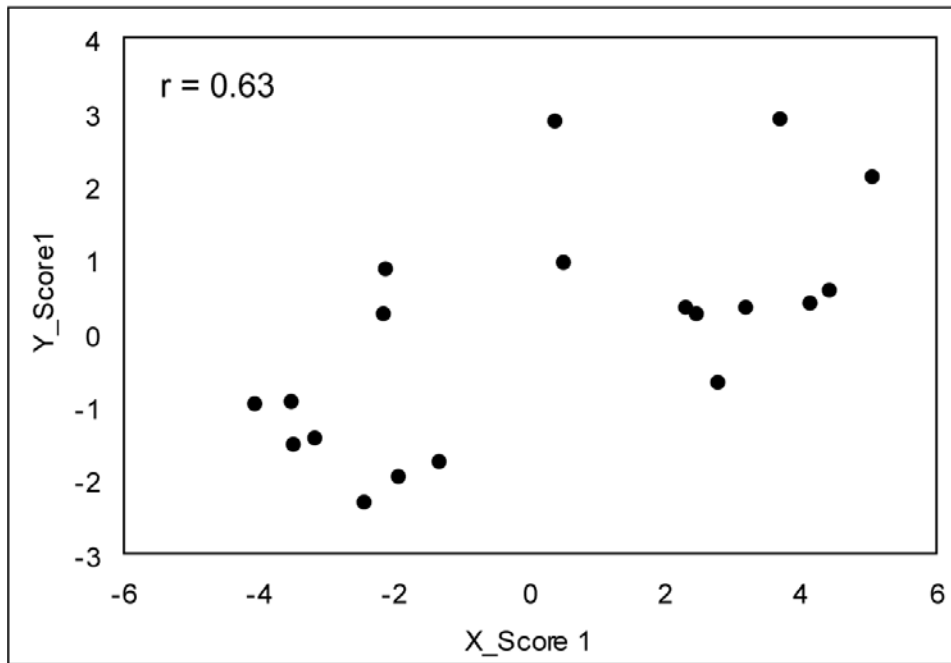


Figure 10. Plot of X_ and Y_Scores for factor 1 for each observation for the PLS model of the 5 biota and 14 landscape variables for the Blue Ridge (BR) ecoregion.

b) **Piedmont (P):** There were three significant factors that accounted for 29 and 86 percent of the variability for the biota and landscape metrics, respectively (Table 10). X_ and Y_Scores and weights for the model are shown in Figures 11 and 12. The slopes, percent of total area with slopes >3 percent (Slope3), and erodible soil (Soil_er) were the most important variables (VIP >1, Table 11). Percent crop and Crop_slp_mod were also important (VIP ≈ 1). Crop, pasture, and agriculture on the slopes were all grouped in the first quadrant of Figure 12 (0.8 < VIP < 1.0); they have positive impacts on factor 1 and on factor 2. Stream density is less important here than in the BR ecoregion. Over half of the P ecoregion has terrain with slopes greater than 3 percent. The most predominant landcover is forest, followed by agriculture (with a nearly equal split between pasture and row crops) and transitional barren areas. Agriculture on slopes greater than 3 percent is evident throughout the ecoregion. All of the highly erodible soils in the basin are in this ecoregion; only very small patches of low-erodibility soils occur in the P, generally along the outer edge of the basin. Percent forest and stream density are generally less than that of the BR but much greater than that of the Coastal Plain. Agricultural land uses are correlated positively with the biological variable AgPT. AgPT is highly correlated with nutrient concentrations and is representative of a short-time interval; i.e., a high AgPT is likely to indicate a recent influx of nutrients into the water body. Runoff of agricultural fertilizer is a likely source of these nutrients.

Table 10. The significant factors of the final PLS model for the 5 biota and 14 landscape metrics and percent variation accounted for by PLS factors for the Piedmont (P) ecoregion (n = 59).

Blocked Cross Validation for the Number of Extracted Factors			
No. of Extracted Factors	Root Mean PRESS	t ²	P > t ²
0	1.0645	9.8990	0.0470
1	1.0076	13.0935	0.0130
2	0.9834	9.1652	0.0460
3	0.9761	0	1.0000
4	0.9988	3.2603	0.7310
5	1.0139	7.1362	0.1760
6	0.9933	5.9928	0.3290
7	1.0201	5.4862	0.3710
8	0.9919	1.9623	0.8760
9	0.9813	2.4957	0.8120
10	0.9878	2.7186	0.7720
11	0.9887	2.5196	0.8070
12	0.9799	2.8799	0.7570

Minimum root mean PRESS		0.9761	
Minimizing number of factors		3	
Smallest number of factors with P>0.1		3	

Percent Variation Accounted for by PLS Factors				
No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	45.7415	45.7415	16.3575	16.3575
2	26.4205	72.1620	8.5655	24.9230
3	13.6948	85.8567	4.2633	29.1863

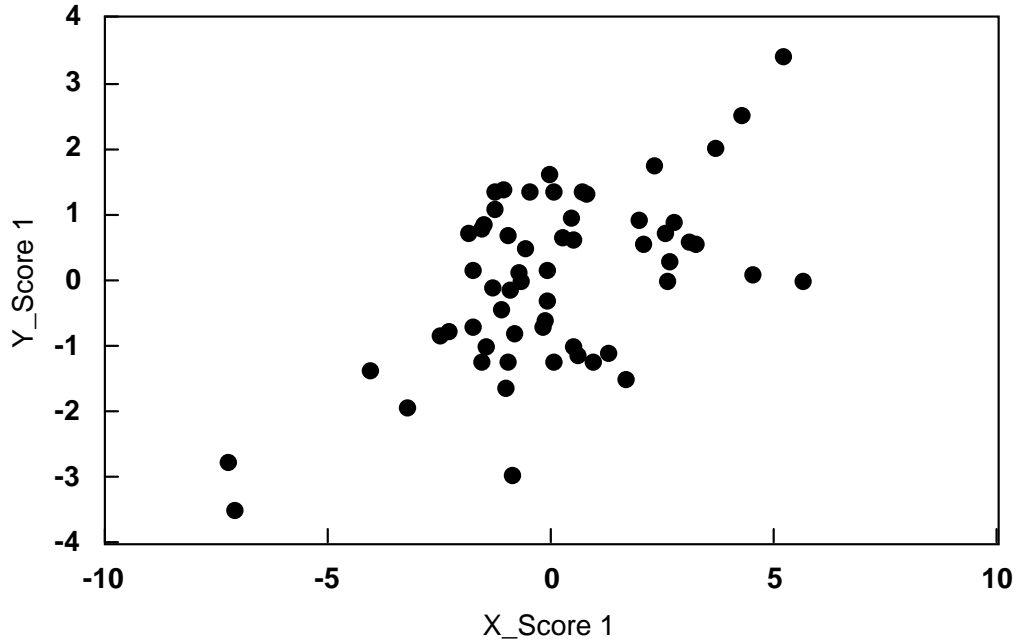


Figure 11. Plot of X_ and Y_Scores for factor 1 for each observation for the PLS model of the 5 biota and 14 landscape variables for the Piedmont (P) ecoregion.

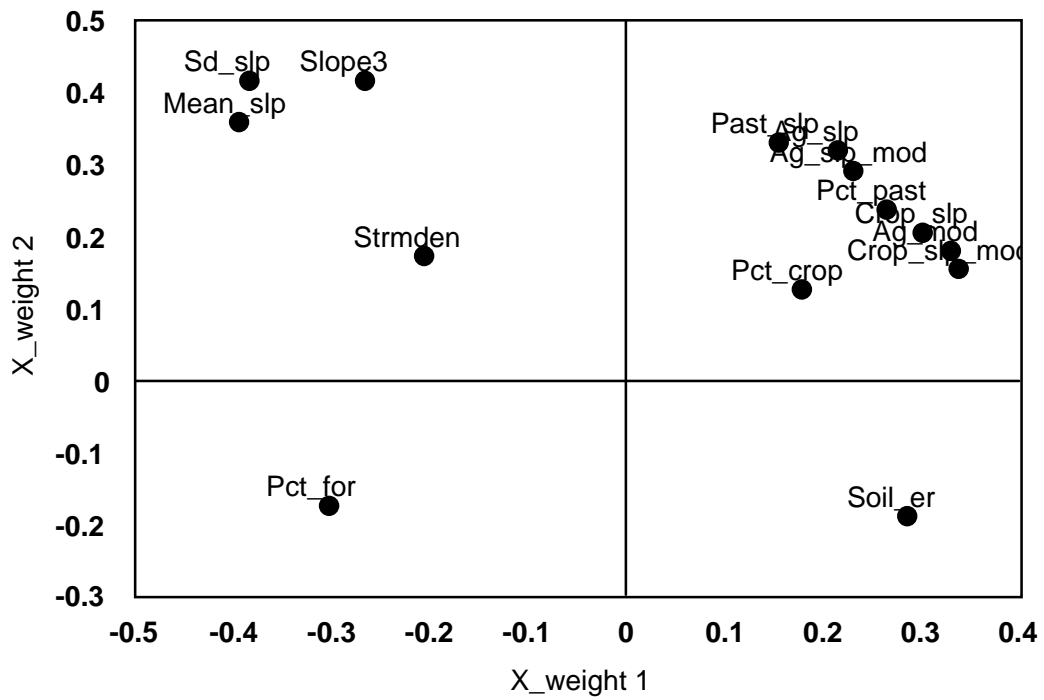


Figure 12. Plot of first and second X_Weights for the PLS model of the 5 biota and 14 landscape metrics for the Piedmont (P) ecoregion.

Table 11. Regression coefficient and VIP values for the Piedmont (P) PLS model.

Predictor	AgPT	EPT	IBI	Hab	Rich	VIP
Ag_mod	0.0658	-0.0531	0.0044	-0.0493	-0.0401	0.9402
Ag_slp	0.0542	0.0220	-0.0230	-0.0336	-0.0009	0.8560
Ag_slp_mod	0.0434	0.0051	-0.0150	-0.0592	-0.0200	0.8607
Crop_slp	0.0700	-0.0340	-0.0034	-0.0301	-0.0232	0.8940
Crop_slp_mod	0.0423	-0.0707	0.0143	-0.0976	-0.0700	0.9799
Past_slp	0.0414	0.0433	-0.0287	-0.0314	0.0087	0.7859
Pct_for	-0.0967	0.0351	0.0060	-0.0251	-0.0017	0.9407
Pct_crop	0.1253	0.0098	-0.0244	0.1474	0.0755	1.0449
Pct_past	0.0625	-0.0145	-0.0098	-0.0310	-0.0154	0.8425
Slope3	-0.0596	0.1888	-0.0661	-0.0376	0.0637	1.2129
Soil_er	-0.0499	-0.1730	0.0708	-0.2148	-0.1715	1.1751
Strmden	-0.0274	0.1133	-0.0390	0.0254	0.0592	0.6498
Sd_slp	-0.0169	0.2475	-0.0922	0.1039	0.1528	1.3098
Mean_slp	-0.0279	0.2325	-0.0842	0.0931	0.1409	1.2607

c) **Coastal Plain (CP):** The scarcity of sampling sites ($n = 7$) in the CP precludes an analysis like those completed for the BR and P areas. Therefore, we only present the percent variation that accounted for two PLS factors in Table 12. X_1 and Y_1 Scores and weights for the model are shown in Figures 13 and 14. Erodible soil and all agriculture/soil/slope-related landscape metrics yielded VIPs greater than 1 (Table 13; Figure 14), while percent forest, percent crop, and percent pasture were the least important ($VIP < 0.8$; Table 12) landscape metrics in the CP. Soils in this ecoregion are generally of low erodibility, and the terrain is much flatter than the other two ecoregions, hence, percent of total area with slopes > 3 percent (Slope3) was not as important as in the BR and P ecoregions. Much of the agriculture is in row crops which are subject to runoff, particularly when located on slopes and/or erodible soils. While agriculture on slopes and/or moderately erodible soils represent a relatively small percentage of the total landcover, these metrics may cause significant impacts on stream biology on a local scale. Although not statistically significant, all landscape metrics have positive regression coefficients with AgPT (Table 13) suggesting, as in the P, these landscape metrics may be indicative of sources of nutrient inputs to streams.

Table 12. Percent variation accounted for by PLS factors for the Coastal Plain (CP) ecoregion. This model is not significant and therefore it did not have the sample validation for the number of extracted factors ($n = 7$).

Percent Variation Accounted for by PLS Factors				
No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	37.8935	37.8935	32.7105	32.7105
2	34.7338	72.6273	24.8842	57.5947

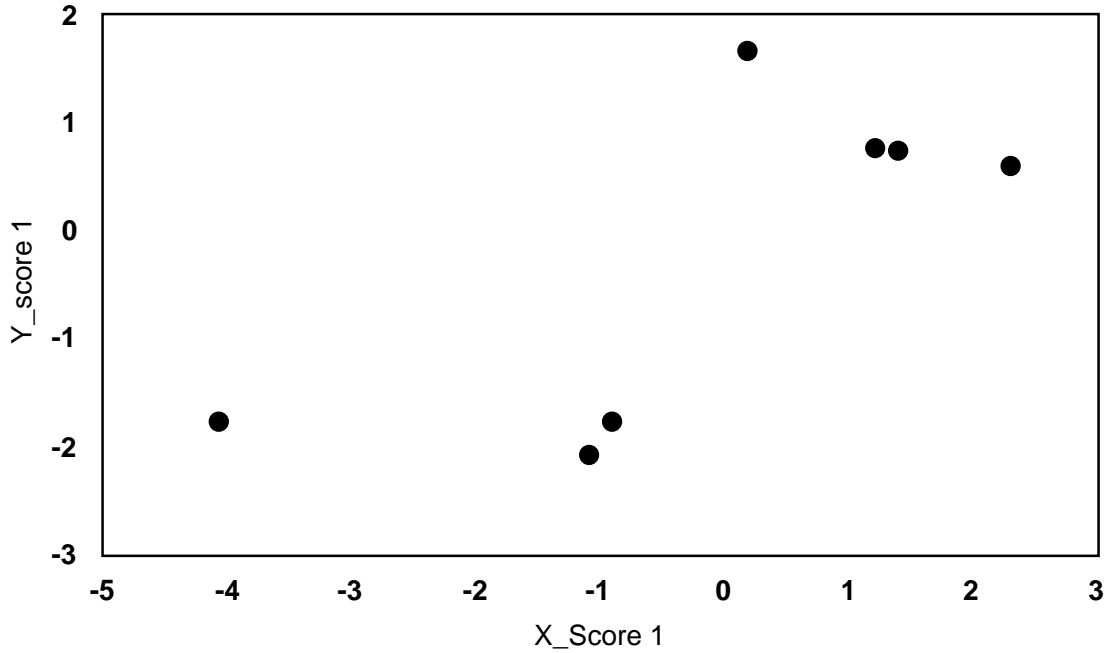


Figure 13. Plot of X_ and Y_Scores for factor 1 for each observation from PLS model of the 5 biota and 14 landscape variables for the Coastal Plain (CP) ecoregion.

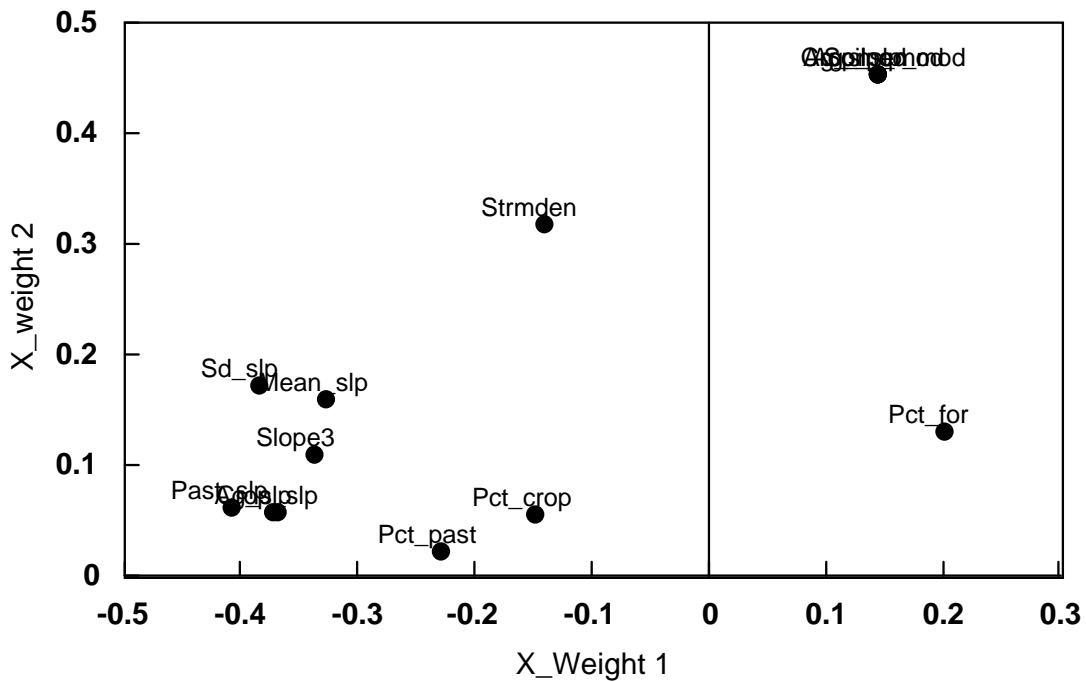


Figure 14. Plot of first and second X_Weights for the PLS model of the 5 biota and 14 landscape metrics for the Coastal Plain (CP) ecoregion. The variables plotted on top of each other in the upper right-hand corner are Soil_er, Crop_slp_mod, Ag_slp_mod, and Ag_mod. Cluster of variables in lower left-hand corner are Past_slp, Crop_slp, and Ag_slp.

Table 13. Regression coefficients of the biota and VIP for landscape metrics for the Coastal Plain (CP) ecoregion.

Predictor	AgPT	EPT	Hab	Rich	VIP
Ag_mod	0.1997	-0.0257	0.0855	0.0150	1.1708
Ag_slp	0.0116	0.1249	-0.1051	0.0903	1.0488
Ag_slp_mod	0.1997	-0.0257	0.0855	0.0150	1.1708
Crop_slp	0.0116	0.1236	-0.1040	0.0894	1.0379
Crop_slp_mod	0.1997	-0.0257	0.0855	0.0150	1.1708
Past_slp	0.0114	0.1366	-0.1154	0.0986	1.1470
Pct_for	0.0646	-0.0595	0.0722	-0.0313	0.6449
Pct_crop	0.0184	0.0513	-0.0388	0.0393	0.4356
Pct_past	0.0012	0.0762	-0.0660	0.0541	0.6410
Slope3	0.0349	0.1155	-0.0896	0.0875	0.9763
Soil_er	0.1997	-0.0257	0.0855	0.0150	1.1708
Strmden	0.1314	0.0610	-0.0115	0.0650	0.8689
Sd_slp	0.0599	0.1336	-0.0974	0.1045	1.1478
Mean_slp	0.0565	0.1146	-0.0819	0.0905	0.9908

Prediction for the Non-Sampled Sites

Prediction for missing or non-sampled measurement(s) for the dependent variables (biota) can be accomplished with PLS. Prediction for missing biota variables is done in two ways; first, the models in “All Data,” above, used all missing and non-missing biota combined (n = 86). As mentioned earlier, PLS has the capability of prediction for the dependent variables if observation(s) are available for the independent variables. The predicted values for the missing biota variables were compared to measured variables by using elevation (as described in Appendix I). Values predicted by the two methods were not significantly different (IBI; $t = 0.25$, $p = 0.80$, $df = 20$; Hab; $t = 0.61$, $p = 0.56$, $df = 8$). Prediction for the missing measurements for the biota using complete data sets that may or may not have missing values in the dependent variables is possible using PLS, an option that is not given by other multivariate analyses.

Another way to predict for dependent variables when they are not measured is to extract a subset of observations with no missing independent variables and use them in PLS model. This model is then used to predict for observations where the dependent variables (e.g., biota) are not measured. For this, we examined the PLS prediction capability using the data for “P” (n = 52) with no missing data and 12 landscape metrics (exclude Sd_slp and Mean_slp). The prediction ability of the model was lower when AgPT was included, therefore, the final model did not include AgPT. Another reason for deleting AgPT was that this variable did not have any missing values. The fitted PLS model contains eight significant factors that account for 40 and 99 percent of the variability in the biota and landscape metrics, respectively (Table 14). In Figure 15, the coefficients for the biota are presented using parallel coordinates superimposed on a bar chart of VIP of the independent variables to enhance an intuitive understanding of the results. The most important landscape metrics in descending order of VIP were percent area with slopes >3 percent, erodible soil, and stream density (Figure 15). Slopes >3 percent enhanced EPT and Rich, but decreased IBI and Hab. Erodeable soil decreased all four measures of biota. EPT responded positively to Slope3 and negatively to erodible soil; other than that it was fairly stable in its relationship to other landscape metrics. Habitat, on the other hand, was positive for stream density, percent forest, percent barren, agriculture on moderately erodible soil, and percent pasture, but responded

negatively for other landscape metrics. IBI responded strongly only for percent forest. Comparing predictions using the PLS models from the P only and all data gave predictions not significantly different (IBI; $t = 0.54$, $p = 0.60$, $df = 20$, Hab; $t = 0.26$, $p = 0.81$, $df = 8$). This is probably related to a well-sampled study area.

Values may be missing in the dependent and/or independent variables. An observation for which an *independent* variable is missing will be deleted from the PLS analysis, and no prediction will occur. An observation of which a *dependent* variable is missing but has no missing values for the independent variable is not included in the analysis, but it will have predicted values. To deal with missing values in the dependent variables, PLS requires more than one dependent variable; and the more dependent variables, the better. The recommendation is ≈ 20 dependent variables with missing values comprising no more than 10-20 percent (Wold, 1995).

In a study area, access to sampling sites and cost of sampling or other reasons may preclude a complete set of samples on the dependent variables, such as water quality properties. Landscape metrics, on the other hand (independent variables), can be obtained for all sites. A PLS model that had been developed for a nearby well-sampled area can be used to predict for the missing dependent variables. We developed a PLS model for the “P” ecoregion using non-missing data and without AgPT. The latter model was used to predict for the missing biota (see Appendix III).

Correlation Between the Dependent Variables

All dependent variables (e.g., biota) may be kept in PLS analysis, even if they are strongly correlated (Wold 1995). But if they are uncorrelated, predictions should be done separately on each of the dependent variables or groups of correlated variables. The strength of collinearity between the dependent variables can be detected by principal component analysis (PCA) using the percentage of variation that each principal component accounts for. When the number of significant principal components is less than the number of dependent variables, it is an indication of collinearity. Unfortunately, there is not a significance test for the number of principal components in SASTM using PRINCOM. Proc FACTOR with option (method = maximum likelihood (ML)) provides a test of significance for the number of factors. Proc FACTOR requires multinormality for the dependent variables and can be tested using %multinorm macro in SASTM (Appendix II). This macro outputs a Q-Q plot of the squared distance and χ^2 , and if the general behavior is on a straight line, then multicollinearity is implied. The biological data were multinormal using the Q-Q plot.

In factor analysis there are two hypotheses that can be tested. The first hypothesis (see below) is to test whether there is no correlation between the original variables. The hypothesis (H_0) of independence between the original biota variables is rejected ($p < 0.0001$). The second hypothesis (H_0) states that 2 factors are adequate to describe the biological data ($p = 0.3545 > 0.05$).

	Test	Df	χ^2	$p > \chi^2$
Hypothesis 1	H_0 : No common factors H_a : At least one common factor	10	96.5216	<.0001
Hypothesis 2	H_0 : 2 factors are sufficient H_a : More factors are needed	1	0.8571	0.3545

Table 14. The significant factors of the final PLS model for the 4 biota and 12 landscape metrics and percent variation accounted for by PLS factors for the Piedmont (P) ecoregion with no missing data (n = 52).

Blocked Cross Validation for the Number of Extracted Factors			
No. of Extracted Factors	Root Mean PRESS	t ²	P > t ²
0	1.0537	11.2336	0.0140
1	1.0572	14.6516	0.0010
2	1.0117	9.1937	0.0350
3	1.0075	8.1123	0.0650
4	1.0550	12.3995	0.0050
5	1.0221	8.3235	0.0610
6	1.0185	7.6355	0.0890
7	1.0424	10.1230	0.0120
8	1.0350	6.6069	0.1610
9	1.0449	9.3328	0.0160
10	1.0364	2.8318	0.6690
11	0.9873	0	1.0000
12	0.9873	3	<0.0001

Minimum root mean PRESS	0.9873
Minimizing number of factors	11
Smallest number of factors with P>0.1	8

No. of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
	1	48.9961	48.8861	10.7059
2	21.0425	70.0386	14.3107	25.0167
3	10.6343	80.6729	4.0113	29.0280
4	5.4136	86.0862	4.0112	33.0395
5	5.5537	91.6401	2.5154	35.5549
6	4.3696	96.0098	1.4616	37.0165
7	2.0516	98.0614	1.2405	38.2571
8	1.2262	99.2876	1.5392	39.7963

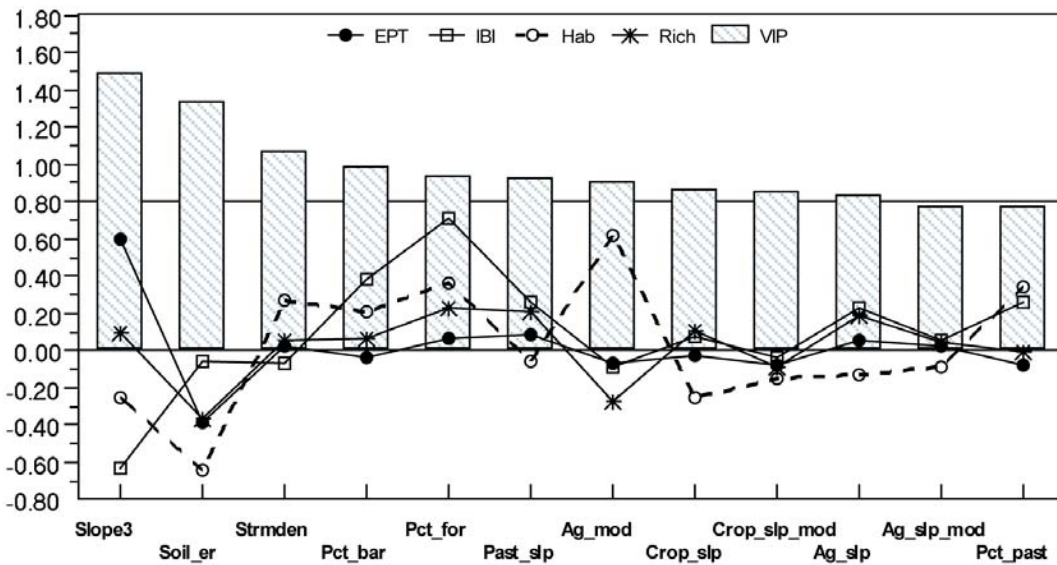


Figure 15. Regression coefficients of landscape metrics for each biota and VIP value for the PLS model for each landscape metrics with no missing biota for the Piedmont (P) ecoregion. Variability explained for landscape and biota was 99 and 40 percent, respectively.

Factor analyses indicated that two factors are needed with squared canonical correlations of 0.93 and 0.57 for factors 1 and 2, respectively. The biota variables of EPT and Rich had high weights in the first factor. We can also utilize PCA to (1) show the weights of EPT and Rich on the second principal components (Figure 16) and (2) to detect any clustering in sites (observations) by plotting the scores of the first two principal components (Figure 17). If clustering is evident, then PLS has to be done separately on each cluster of observations. Our data did not show any clustering pattern (Figure 17). From PCA and factor analyses, we concluded that the dependent variables (biota) were correlated and therefore, all were included in PLS.

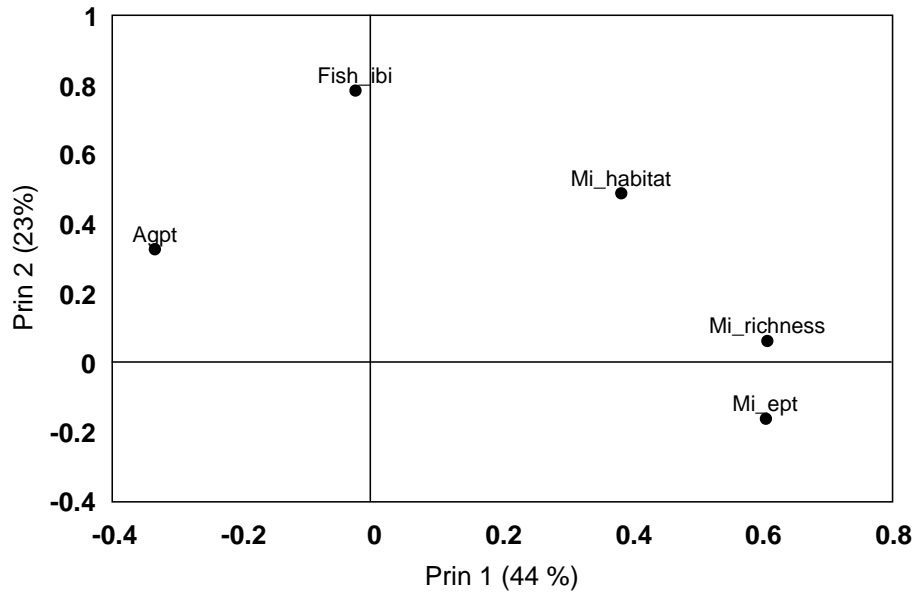


Figure 16. Loading of each of the biota variables (dependent) on the first two principal components (Prin 1 and Prin 2).

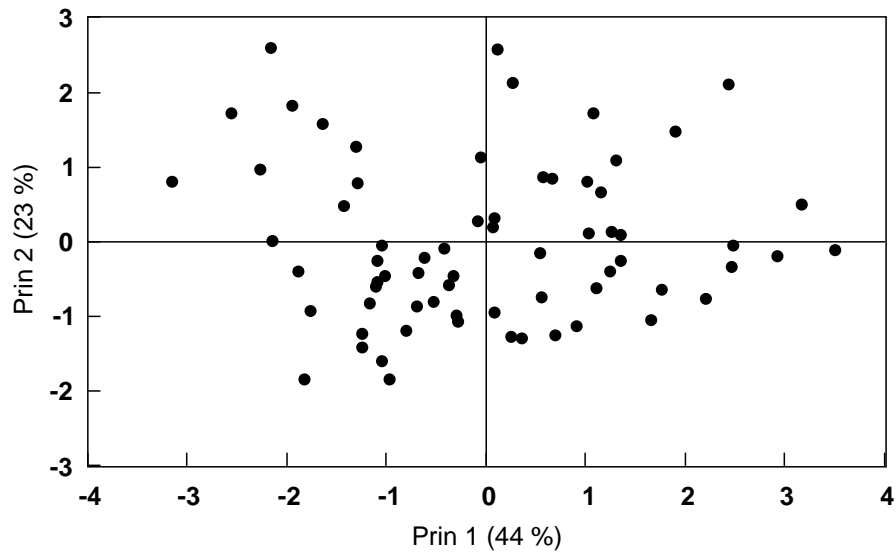


Figure 17. Scores for principal components (Prin 1 versus Prin 2) for the biota variables showing no cluster pattern in sites.

Collinearity is a big concern in multiple regression (see text above), but it is not a concern in PLS. Over fitting the model and consequently having less predictive power than when collinearity exists is still a concern in PLS. PLS overcomes this problem by extracting only the significant factor(s) for the final model, thereby preventing over fitting of the model.

In multiple regression, it is important to have enough observations compared to the number of independent variables. The predictive ability is low when the number of independent variables is large in comparison to the number of observations. In contrast, PLS can predict for the dependent variables, even if the above condition exists. PLS has the ability to screen for the most important variables that contribute the most to the variation in the dependent variables. Hence, a final model can be obtained with the variables contributing the most to describe a biological phenomenon.

Summary

PLS permitted analyses of the data by ecoregion even for relatively small sample sizes, an option that is not available with other multivariate analyses (e.g., canonical correlation). The analyses revealed that different landscape metrics affect surface water biota for different ecoregions. Percent forest (Pct_for) and percent of total area with slopes >3 percent (Slope3) were the most important landscape variables in Blue Ridge; percent of total area with slopes >3 percent, mean and standard deviations of slope, soil erodibility, percent crop, and crops on slopes >3 percent with moderately erodible soils (Crop_slp_mod) were the most important in the Piedmont; and soil erodibility, agriculture on slopes >3 percent and on moderately erodible soils (Ag_slp_mod), and pasture on slopes were the most important landscape variables in Coastal Plain. Stream density was more important in the Coastal Plain than other ecoregions. Erodeable soil was the common landscape variable in the Piedmont and Coastal Plain. This suggests that management practices might target areas with erodible soils in these ecoregions.

When a prediction for missing dependent variables is desired, developing a model from data with no missing measurements for the dependent variables is preferable to PLS. Model performance was best for the Piedmont ecoregion, and this model was used to predict the biota from the elevation metric in other locations. Although model differences between prediction of Piedmont separately and all data combined were insignificant, we recommend using the model that explains the greatest amount of variability in the data.

5.2.3.2 Application of PLS to Other Ecological Data

Western Lake Erie Wetland

We wanted to demonstrate the utility of PLS to other ecological data in prediction of responses. We used data from a Western Lake Erie wetland (Lopez et al., 2004). Assemblages of common reed (*Phragmites australis* (Cav.) Steudel) tend to form large, relatively monospecific, “stands” that may predominate in wetland plant communities, supplanting other plant taxa (Marks et al., 1994). Compared to other more heterogeneous plant communities, *Phragmites* stands are less suitable as animal habitat and reduce the overall biological diversity of wetlands. Within patches of *P. australis*, the percent of vegetational cover, stem density, and plant height are important metrics for assessing and monitoring the influx and extent of this opportunistic wetland plant species (Lopez et al., manuscript in preparation).

Vegetational cover of *P. australis* can be predicted and mapped using advanced remote sensing techniques and geographic information systems (Lopez et al., 2004). A similar direct relationship between the reflectance values derived from remote sensing data and either stem density or plant height cannot be applied. However, plant height and stem density may relate to cover metrics in *P. australis*. We leveraged this multivariate relationship to predict *P. australis* height and stem density using field-sampled plant structural characteristics and PLS.

Our objective was to examine the utility of PLS for estimating *P. australis* height and stem density solely using cover measurements. Successful use of PLS to estimate plant height and stem density in this manner has important implications for measurement of these plant characteristics from remote sensing platforms, but it is beyond the scope of this report. The field-sample variables used in the analyses are described in Table 15 (n = 221 sites).

Table 15. Variables used in the analyses with a description of each field-sampled plant structural characteristic. Variables were measured in either 1 m² or 3 m² quadrats, as indicated. Target species = *P. australis*; nontarget species = all remaining plant species observed within quadrats; understory = an approximate field measurement of vegetation from the soil surface to 0.5 m; canopy = an approximate field measurement above a height of 0.5 m; and D1-D5 = diameters of 5 plants located at the corner and the center of 1 m² quadrant.

Variable Name	Description
Ht	Mean height of target species (m) in canopy (3 m ²)
NumL	Number of live target species stems (1 m ²)
%LTC	Percent cover live target species in canopy (3 m ²)
%LNTC	Percent cover live nontarget species in canopy (3 m ²)
%totTC	Percent total target species in canopy (3 m ²)
%totNTC	Percent total nontarget species in canopy (3 m ²)
%totNTU	Percent total nontarget species in understory (3 m ²)
%LNTU	Percent cover live nontarget species in understory (3 m ²)
D1-D5	Stem diameter (mm) of five <i>P. australis</i> (1 m ²)

Results

With 11 independent (predictor) plant community variables (%LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5) and two dependent (response) variables (ht and numL), the PLS model yielded three significant factors (Table 16) that explained 65 and 72 percent of the variability in plant height and the live stem density and the remaining plant community variables, respectively. Model predictive ability, as measured by three significant factors, was higher than that of the null (0 factor) by 39 percent (PRESS; null = 1.03; factor 3 = 0.63). Bi-plots of weights for plant community variables indicate which variables have more influence in the two linear compositions of the variables from the two data sets on a PLS factor (Figure 18). It is apparent that, for example, percent total target species in canopy (%totTC) and percent live target species in canopy (%LTC) are projected close to each other, and their contributions are more than that of measurements for the nontarget species. Since there are no variables that are clustered near the origin (i.e., have low weights for both factors), all variables contributed substantially to the predictive capability of the model. Those variables (e.g., 5 plant diameters) that cluster near each other indicate equal weight on a factor. Bi-plots of the scores for ht and numL and %LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5 for the first PLS factor indicate the strength of relation between the two linear compositions ($r = 0.73$; Figure 19). The VIP, which accounts for weight values, represents the relative importance of each variable in the PLS model for both the dependent (ht and numL) and independent variables. The importance (as determined by the VIP) of each plant community variable is high ($VIP \geq 0.8$; Table 17). Percent live nontarget species in canopy (%LNTC), percent total nontarget species in canopy (%totNTC), percent total nontarget species understory (%totNTU), and percent live cover target species understory (%LNTU) were contributed

equally with VIP close to 0.8. Percent target species in canopy (%LTC) and percent total target species in canopy (%totTC) contributed the most to the ht and numL (VIP ≥ 1).

Table 16. The significant factors for the PLS model for the numL, ht and 11 plant community variables (%LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5). Split-sample validation for the number of extracted factors was used.

Number of Extracted Factors	Root Mean Press	t ²	P > t ²
0	1.0270	46.168	<0.0001
1	0.7373	20.4217	<0.0001
2	0.6678	11.9253	0.0030
3	0.6297	0	1.0000
4	0.6378	5.1253	0.0600
5	0.6317	0.2327	0.0600
6	0.6355	1.1388	0.5670
7	0.6386	1.1645	0.7030
8	0.6411	1.3108	0.6320
9	0.6423	2.3962	0.2840

Minimum root mean PRESS	0.6297
Minimizing number of factors	3
Smallest number of factors with P > 0.05	3

Percent Variation Accounted for by PLS Factors				
Number of Extracted Factors	Model Effect		Dependent Variables	
	Current	Total	Current	Total
1	49.8306	49.8306	49.1162	49.1162
2	14.4155	64.2461	10.2431	59.3592
3	7.5968	71.8429	5.4486	64.8078

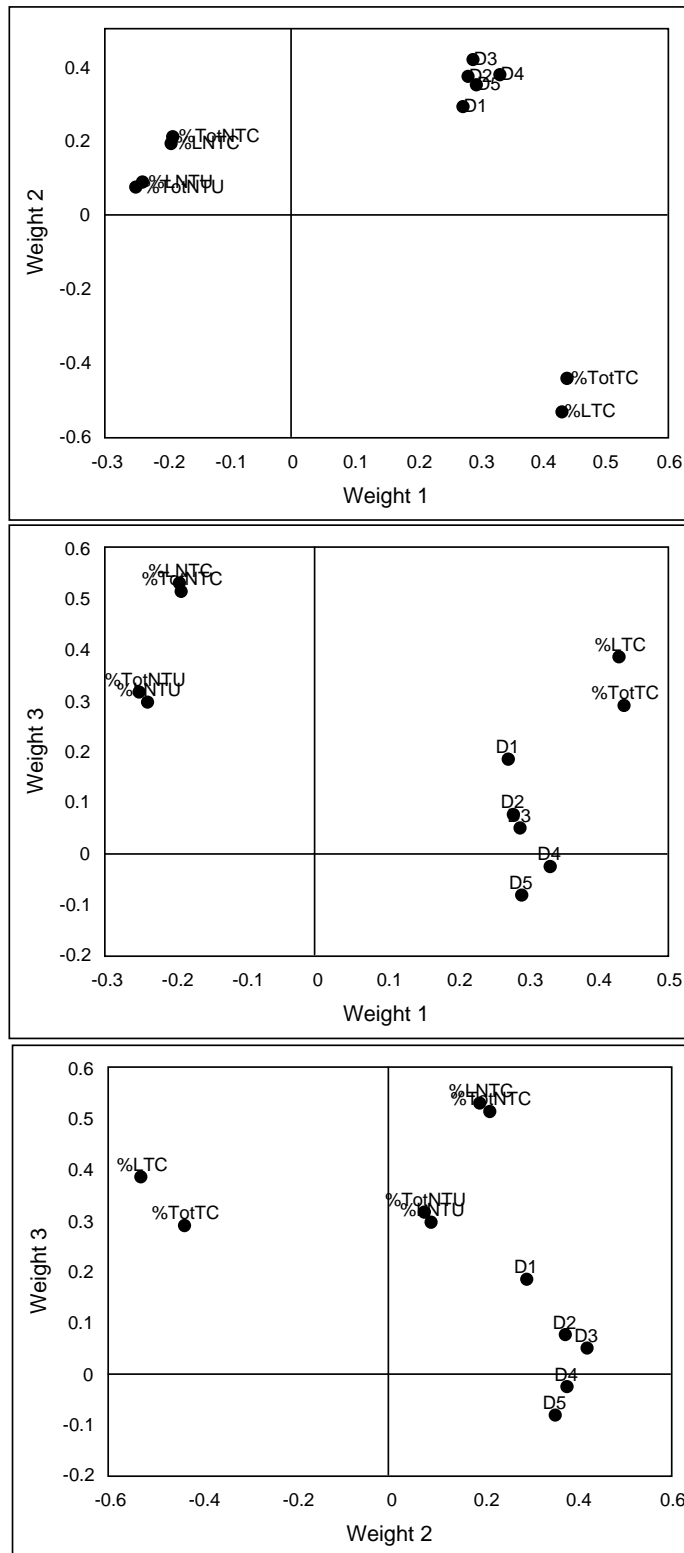


Figure 18. Linear compositions of the predictors (%LNTC, %LTC, %totNTC, %totNTU, %totTC, %LNTU, and D1-D5) of plant community for the three significant PLS factors.

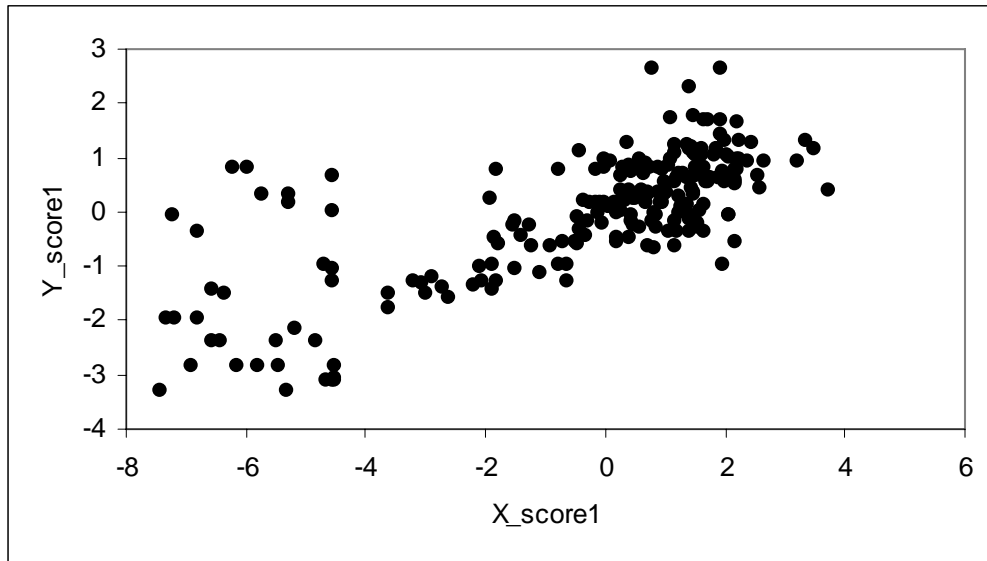


Figure 19. The relation between *P. australis* height/number and percent cover scores for the first PLS factor ($R^2 = 0.53$). X_score1 is linear composition of %LNTC, %LTC, %totNTC, %totNTU, %totTC, %LNTU, and D1-D5 for the first significant factor; Y_score1 is the linear composition of number and height of the target plant for the first significant factor.

Table 17. Coefficient values for the 2 responses (ht and numL) and 11 predictors (%LNTC, %LTC, %totNTC, %totTC, %totNTU, %LNTU, and D1-D5) and VIP for the 11 plant metrics for the PLS model.

Predictor	ht	NumL	VIP
%LNTC	-0.005	0.031	0.800
%LTC	0.134	0.532	1.478
%totNTC	-0.005	0.016	0.793
%totNTU	-0.053	-0.003	0.790
%totTC	0.136	0.456	1.421
%LNTU	-0.049	-0.013	0.755

Predictor	ht	NumL	VIP
D1	0.133	0.006	0.896
D2	0.134	-0.070	0.955
D3	0.139	-0.101	1.007
D4	0.144	-0.094	1.086
D5	0.123	-0.113	0.973

To examine the data for outliers, relationships between observed and predicted values for each of the responses (number and height) and predictors (percent cover data set) are plotted by Euclidean distance (Figure 20). Any point (or group of points) with distinctive and large deviation(s) from the general distribution of the data was examined. There is no evidence of serious outliers in the data.

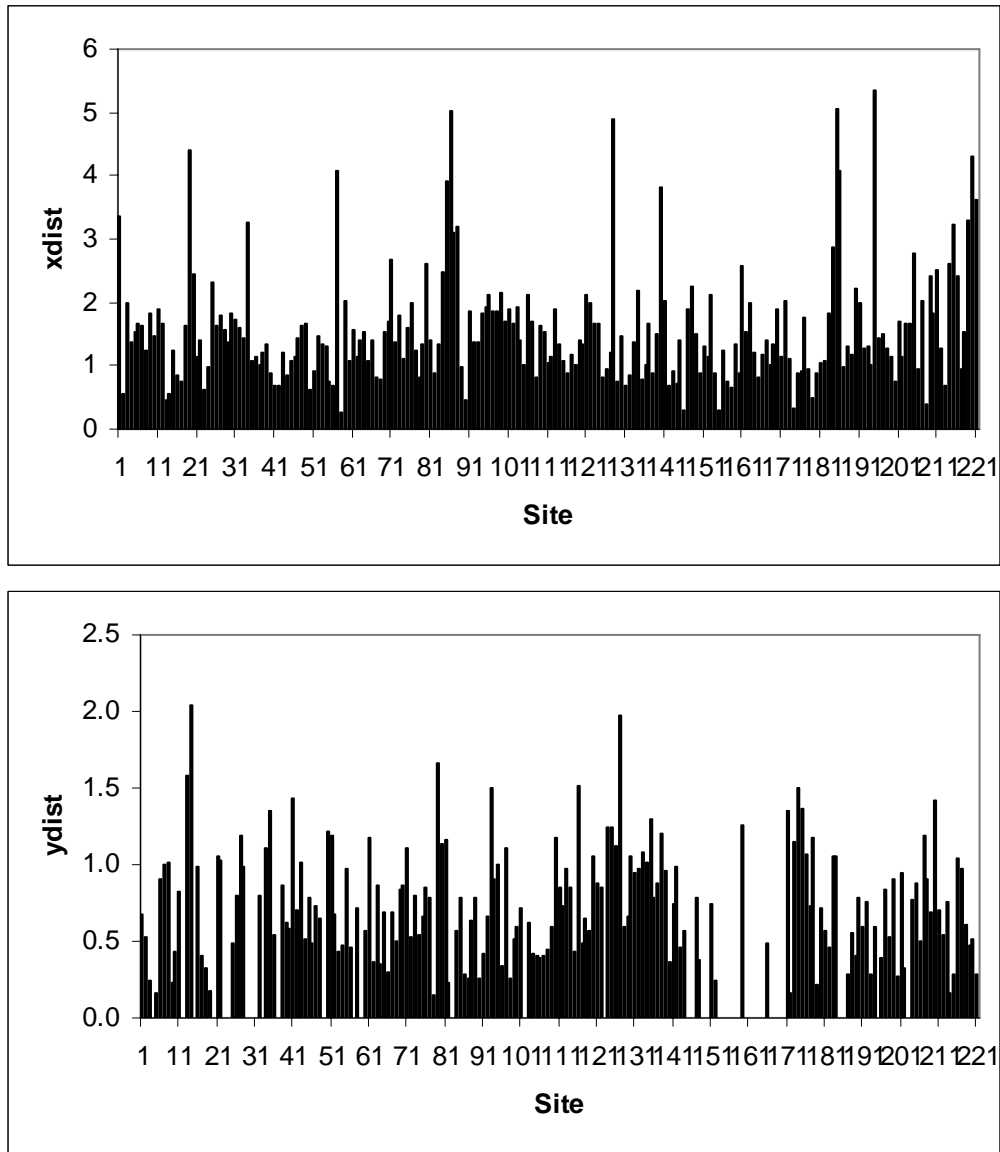


Figure 20. Final PLS model residuals: (above) distance from X_variables to the final PLS model and (below) distance from Y_variables to the final PLS model.

Conclusion

These results suggest that percent cover can be predictive of the height and stem density of *P. australis*. These results indicate that percent cover, as measured by remote sensing reflectance data, may be an indicator of plant size and density, which are both important measures of the predominance of this opportunistic plant in wetlands.

Section 6

Comparison with Canonical Correlation

Two multivariate analyses, canonical correlation and partial least square (PLS) regression, were conducted to study the relationships between landscape metrics and surface water quality. Canonical correlation is well known in biological and ecological studies, but to our knowledge, this is the most current use of PLS to explore relationships in landscape studies. Although PLS is unfamiliar to many statisticians, it has been used extensively in chemistry to describe relationships between chemical structures and activity. We wanted to test PLS as a potentially parsimonious method for landscape ecologists, faced with issues of collinearity and small sample size, to explore relationships which may be used to assess the quality and vulnerability of an ecosystem. We ran PLS only for the biota landscape data to permit detailed exploration of the relationships.

In canonical correlation analyses, collinearity, missing observations, multinormality, and the ratio of number of variables to number of observations are important issues that need to be dealt with prior to analysis. PLS is less subject to these constraints. For the canonical correlation analyses, landscape metrics were selected based on pairwise correlation and discriminant analyses resulting in a total of four landscape metrics. For the PLS analyses, all landscape variables (26) were initially considered in the model and the most important ones were included based on their VIP (>0.8) in the final model. VIP provides information not only as to how important each landscape variable is (e.g., Table 7) but how similar the contribution is to that of other variables. For example, landscape metrics such as Crop_slp_mod, Ag_slp_mod, and Ag_mod had the same VIP as Soil_er (Table 13) indicating equal contributions of soil erodibility and crop/agriculture on slopes with moderately erodible soil in predicting biota metrics. Therefore, for this group of landscape variables, one variable (e.g., soil erodibility) alone may be used in any model.

The percent of total area with slopes >3 percent (Slope3) had the largest canonical coefficient in the land-biota analysis (page 20) and had a large VIP in PLS. The only other landscape variable common to canonical correlation and the initial PLS model was Pct_bar, which was not greatly important in either method. Pct_bar did not make it to the final model because of its low VIP value. In contrast to the canonical model, the PLS model indicated that landscape metrics such as percent forest and agriculture on moderately erodible soil were the second most important landscape variable group. The other intriguing feature, PLS allowed the analyses by ecoregion, which revealed different landscape metrics related to surface water biota based on their spatial association (ecoregion, Table 18). It is evident that the importance of any landscape metric is a function of its spatial location in the study area. The importance of percent forest and percent pasture was the highest in BR and decreased consistently across the adjacent ecoregions of the study area; the opposite relationship was found for Crop_slp and soil erodibility. Percent crop, percent of areas on slopes >3 percent, and soil erodibility are the most important landscape variables in the Piedmont. The relative importance of percent pasture on slopes >3 percent increased in the Blue Ridge and Coastal Plain.

Table 18. Rank of the landscape metrics in the PLS model using VIP levels. "All" is the overall model for the three ecoregions.

Metrics	All	BR	P	CP
Ag_mod	1	1	2	1
Ag_slp	2	1	2	1
Ag_slp_mod	2	1	2	1
Crop_slp	2	3	2	1
Crop_slp_mod	2	3	2	1
Past_slp	2	1	3	1
Pct_for	1	1	2	3
Pct_crop	3	2	1	3
Pct_past	2	1	2	3
Slope3	2	1	1	2
Soil_er	3	3	1	1
Strmden	3	3	3	2
Sd_slp	1	1	1	1
Mean_slp	1	1	1	2

1 = VIP >1

2 = VIP between 0.8 and 1

3 = VIP <0.8

References

- Barcikowski, R.S. and J.P. Stevens. 1975. A monte carlo study of the stability of canonical correlation, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research*. 10, 353-364.
- Chaloud, D.J., C.M. Edmond, and D.T. Heggem. 2001. Savannah River Basin Landscape Analysis. EPA/600/R-01/069.
- Clark, D. 1975. Understanding canonical correlation analysis. Geo Abstract Ltd. University of East Anglia, Norwich, NR4 7TJ.
- Geweke, J.F. and K.J. Singleton. 1980. Interpreting the likelihood of ratio statistics in Factor models when sample size is small. *Journal of the American Statistical Association*. 75:133-137.
- Gittins, R. 1985. Canonical analysis: A review with application in ecology. Springer Verlag, New York.
- Griffith, D.A. and C.G. Amrhein 1997. *Multivariate Statistical Analysis for Geographers*. Prentice Hall, New Jersey.
- Hair, J.F., R.E. Anderson, R.L. Tatham. 1987. *Multivariate data analysis with readings*. Macmillan Pub. Co., New York.
- Johnson, K.W. and N. Altman. 1996. Canonical correspondence as an approximation to Gaussian Ordination. Technical Report BU-1349-M. Cornell University, Ithaca, New York.
- Johnson, R.A. and Wichern, D.W. 2002. *Applied multivariate statistical analysis*. Prentice Hall, New Jersey.
- Jones, B.K., A.C. Neale, M.S. Nash, R.D. Van Remortel, J.D. Wickham, K.H. Riitters, R.V. O'Neill. 2001. Predicting Nutrient and Sediment Loadings to Streams from Landscape Metrics: A Multiple Watersheds Study from the United States Mid-Atlantic Region. *Landscape Ecology*. 16(4) 301-312.
- Karr, J.R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6(6): 21-27.
- Lawley, D.N. and A.E. Maxwell. 1971. *Factor Analysis as a statistical method*. Macmillan Publishing Co., Inc., New York.
- Lenat, D.R. 1987. "The Macroinvertebrate Fauna of the Little River, North Carolina: Taxa List and Seasonal Trends," *Arch. Hydrobiol.* 110(1), 19-43.
- Lopez, R.D., C.M. Edmonds, D.T. Heggem, A.C. Neale, K.B. Jones, E.T. Slonecker, E. Jaworski, D. Garofalo, and D. Williams. 2004. Accuracy Assessments of Airborne Hyperspectral Data for

- Mapping Opportunistic Plant Species in Freshwater Coastal Wetlands *in* (R.S. Lunetta and J.G. Lyon, eds.) *Remote Sensing and GIS Accuracy Assessment*. CRC Press, Boca Raton, FL. 304pp.
- Madansky, A. 1988. Prescriptions for working statisticians. Spring Verlag, New York.
- Marks, M., B. Lapin, and J. Randall. 1994. *Phragmites australis* (*P. communis*): threats, management, and monitoring, *Nat. Areas J.*, 14, 285-294.
- Mehaffey, M.H., M.S. Nash, T.G. Wade, C.M. Edmond, D.W. Ebert, K.B. Jones, and A Roger. 2001. A Landscape Assessment of the Catskill/Delaware Watershed. EPA/600/R-01/075.
- Nash, M.S. and D. Bradford. 2001. Parametric and Non-parametric (MARS; Multivariate Additive Regression Splines) Logistic Regressions for Prediction of a Dichotomous Response Variable with an Example for Presence/Absence of an Amphibian. EPA/600/R-01/081, 54p.
- Noy-Meir, I. 1974. Multivariate analysis of the semiarid vegetation in southeastern Australia. II Vegetation Catenae and environmental gradients. *Aust J. Bot* 22, 115-140.
- Owen, D.B. 1988. The starship. *Cummun. Statist. -Simula.*, 17(2):315-323.
- Rencher, A.C. 1998. Multivariate statistical inferences and applications. John Wiley and Sons' Inc., New York.
- Ter Braak, C.J.F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* 69: 69-77.
- Thorndike, R.M. 1978. Correlation procedure for research. Gardner Press, New York.
- van der Voet, H. 1994. Comparing the predictive accuracy of models using a simple randomization test. *Chemometric and Intelligent Laboratory Systems* 25: 313-323.
- Wold, S. 1995. PLS for multivariate Linear Modeling *in* Chemometric methods in molecular design methods and principles in medicinal chemistry, (ed) H. van de Waterbeemd, Weinheim, Germany: Verlag-Chemie.

Appendix I

Supplementary Information

Missing Observations

The total number of sites sampled for water quality parameters was 89, with 18 missing values for at least one metric. When statistical correlation is performed, only sites with complete data are used so that 71 observations would be included. IBI was the most frequently missing biological metric with 14 sites missing data, 7 of which were in the Coastal Plain (CP). Because of this, CP would be excluded from the analysis, hence there would be a loss of an ecoregion despite the existence of other water biological metrics. Therefore, we substituted for the missing value of a metric its predicted value from a regression model since that metric regressed to a surrogate variable not in the canonical correlation analysis. Elevation is a suitable surrogate because it is not included in the canonical correlation analysis and has no missing data.

For each water parameter per ecoregion, we studied its behavior with elevation to observe the commonality in behavior. When missing value(s) occurred only in a specific ecoregion, a regression model was fitted only to that ecoregion data. The IBI had missing data in both the CP and Piedmont (P) ecoregions; therefore, we used all non-missing values to cover a wider range for better prediction. The model was chosen based on the significance level of the model-F, R^2 , and coefficient-t values ($p \leq 0.05$). When no significant relationship was found, the average value of that water quality parameter in the ecoregion was used as a substitute for the missing values. The water quality data for the best fit model is given in Table I-1.

Diagnostic checking of the residuals of the best fit model was performed to test the model assumption concerning the residual for independence and normality (Madansky, 1988).

Assumptions

In relating stream biology and landscape metrics, we started with nine landscape metrics (Strmden, Slope3, Soil_er, Pct_past, Pct_for, Pct_bar, Past_slp, Crop_slp_mod, and Ag_mod) and five biological variables (AgPT, IBI, Hab, Rich, and EPT). For the canonical correlation, we had two matrices: stream biota and landscape data sets. Before running the canonical analysis, we tested the data for collinearity, normality, outliers, and the ratio of the number of variables to the number of observations in a sample.

1 - Collinearity: The absolute value of the correlation between variables (within and between each data matrix) was studied (Table I-2). A value of more than 0.9 was considered a sign of collinearity (Griffith and Amrhein, 1997). Therefore, we extracted variables with an absolute value of correlation of less than 0.9. There is a detailed discussion on collinearity in Nash and Bradford (2001).

2 - Normality: Multinormality may not be important when the purpose is only to describe relationships. When testing hypotheses and inferring results, multinormality (joint distribution in canonical) of the

variables is important and has to be met (Gittins, 1980). We used a macro (%multinorm; Appendix II) in SAS™ to test for multinormality of the two data sets. The %multinorm produces a chi-square/Q-Q plot, which describes the relationships between squared distance and chi-square quantile (Figure I-1). The correlation between squared distance and chi-square can be calculated and used to test the hypothesis of normality (Johnson and Wechern, 2002; Table 4.2, pg. 182). For the landscape and biological data sets, the correlation between the squared distance and chi-square is 0.9905 (n = 84), which is greater than the tabulated value (0.9822, n = 100, $\alpha = 0.01$) (Johnson and Wechern; Table 4.2) and normality is not rejected. If the relationship between squared distance and chi-square is linear, it is an indication of a symmetrical distribution without long tails, and multinormality is reached. Accordingly, if an observation is extreme and is far from the linear behavior, then that observation has to be studied and deleted, if necessary, to preserve multinormality.

Table I-1. The best fit model for the surface water quality variables with their significant levels. Numbers in parentheses are the total number of missing values.

Variable	Ecoregion	Model	P > F	R ²
EC	BR(2)	avg. = 90.89		
	P(5)	avg. = 92.92 ^a		
	CP(1)	avg. = 65.85		
pH	BR(2)	7.25 - 0.0005 * elev. ^b	0.007	0.37
	P(1)	avg. = 6.92		
	CP(1)	avg. = 6.05		
DO	BR(6)	avg. = 8.33		
	P(3)	avg. = 7.13		
	CP(1)	avg. = 6.63		
IBI	CP(7) & P(6)	- 0.02531 * elev. + 1.69* elev. ^{0.5}	0.014	0.92
Rich	P(1)	avg. = 19.62		
Hab	BR(2)	avg. = 75.89		
	P(2)	avg. = 75.92		
	CP(1)	77.46 + 0.05 * elev.	0.0372	0.70
EPT	P(1)	157.23 * (elev.) ^{-0.5}	<0.0001	0.83

^a Two values were very high (3,260 and 914) and were excluded from the average value.

^b pH = 6.5 at elevation of 109 m was not in the fitted model (outlier).

Table I-2. Pairwise correlation between all variables in the biological and landscape data.

	Variable	AgPT	EPT	IBI	Hab	Rich	Ag_hi	Ag_slp_hi	Ag_mod	Ag_slp	Ag_slp_mod	Bar_slp_hi	Bar_slp_mod	Crop_slp	Crop_slp_mod	Past_slp
	n =	86	85	72	81	85	86	86	86	86	86	86	86	86	86	86
1	AgPT	1.000	-0.338	0.109	-0.097	-0.362	0.385	0.393	0.380	0.376	0.309	0.002	-0.120	0.380	0.318	0.316
2	EPT	-0.338	1.000	-0.175	0.197	0.821	-0.080	-0.080	-0.247	-0.205	-0.168	0.006	0.010	-0.281	-0.220	-0.138
3	IBI	0.109	-0.175	1.000	0.150	0.048	0.177	0.181	0.101	0.154	0.146	0.054	0.059	0.077	0.086	0.170
4	Hab	-0.097	0.197	0.150	1.000	0.275	-0.177	-0.175	-0.283	-0.313	-0.329	-0.033	0.000	-0.205	-0.326	-0.314
5	Rich	-0.362	0.821	0.048	0.275	1.000	-0.046	-0.041	-0.207	-0.184	-0.151	0.097	-0.008	-0.292	-0.232	-0.106
1	Ag_hi	0.385	-0.080	0.177	-0.177	-0.046	1.000	0.998	-0.153	-0.037	-0.153	0.428	-0.062	0.006	-0.157	-0.050
2	Ag_slp_hi	0.393	-0.080	0.181	-0.175	-0.041	0.998	1.000	-0.148	-0.031	-0.148	0.426	-0.060	0.012	-0.152	-0.046
3	Ag_mod	0.380	-0.247	0.101	-0.283	-0.207	-0.153	-0.148	1.000	0.879	0.924	-0.175	-0.211	0.651	0.889	0.849
4	Ag_slp	0.376	-0.205	0.154	-0.313	-0.184	-0.037	-0.031	0.879	1.000	0.964	-0.137	-0.241	0.763	0.816	0.955
5	Ag_slp_mod	0.309	-0.168	0.146	-0.329	-0.151	-0.153	-0.148	0.924	0.964	1.000	-0.170	-0.200	0.634	0.862	0.967
6	Bar_slp_hi	0.002	0.006	0.054	-0.033	0.097	0.428	0.426	-0.175	-0.137	-0.170	1.000	-0.072	-0.133	-0.180	-0.118
7	Bar_slp_mod	-0.120	0.010	0.059	0.000	-0.008	-0.062	-0.060	-0.211	-0.241	-0.200	-0.072	1.000	-0.168	-0.100	-0.237
8	Crop_slp	0.380	-0.281	0.077	-0.205	-0.292	0.006	0.012	0.651	0.763	0.634	-0.133	-0.168	1.000	0.765	0.537
9	Crop_slp_mod	0.318	-0.220	0.086	-0.326	-0.232	-0.157	-0.152	0.889	0.816	0.862	-0.180	-0.100	0.765	1.000	0.713
10	Past_slp	0.316	-0.138	0.170	-0.314	-0.106	-0.050	-0.046	0.849	0.955	0.967	-0.118	-0.237	0.537	0.713	1.000
11	Pct_bar	-0.130	-0.127	0.107	0.230	-0.064	0.020	0.023	-0.451	-0.354	-0.448	0.111	0.630	-0.071	-0.386	-0.430
12	Pct_crop	0.269	-0.338	-0.011	0.118	-0.180	-0.026	-0.022	0.243	0.254	0.131	-0.129	-0.223	0.572	0.269	0.068
13	Pct_for	-0.414	0.488	-0.023	0.135	0.339	-0.031	-0.041	-0.586	-0.629	-0.532	0.132	0.161	-0.670	-0.507	-0.513
14	Pct_past	0.411	-0.221	0.161	-0.291	-0.172	0.000	0.006	0.936	0.920	0.928	-0.106	-0.253	0.576	0.746	0.936
15	Pct_urb	0.133	-0.145	-0.263	-0.322	-0.210	0.138	0.143	0.033	0.060	0.056	-0.022	-0.105	0.041	0.066	0.059
16	Pct_wet	-0.088	-0.250	0.023	0.210	-0.100	-0.053	-0.052	-0.220	-0.182	-0.210	-0.057	-0.067	-0.101	-0.217	-0.191
17	Pct_wtr	0.150	-0.194	0.119	-0.061	-0.164	-0.072	-0.069	0.326	0.298	0.285	-0.112	0.114	0.253	0.256	0.272
18	PwrPlpTI	0.059	0.001	-0.069	-0.105	-0.031	-0.023	-0.021	0.148	0.209	0.164	-0.041	-0.105	0.395	0.300	0.091
19	Slope3	-0.138	0.577	-0.125	-0.131	0.268	-0.092	-0.085	0.015	0.126	0.187	-0.054	0.115	-0.070	0.109	0.197
20	Slp_mod	-0.083	0.290	-0.024	-0.286	0.115	-0.270	-0.261	0.297	0.275	0.412	-0.300	0.197	-0.015	0.370	0.366
21	Sd_slp	-0.178	0.697	-0.085	0.192	0.445	-0.101	-0.096	-0.177	-0.151	-0.100	-0.097	-0.028	-0.252	-0.144	-0.081
22	Mean_slp	-0.210	0.676	-0.108	0.191	0.417	-0.090	-0.086	-0.221	-0.193	-0.136	-0.083	-0.006	-0.301	-0.181	-0.114
23	Soit_er	0.098	0.005	0.143	-0.459	-0.035	0.100	0.097	0.431	0.276	0.414	0.101	0.170	-0.004	0.438	0.362
24	Strmden	-0.263	0.292	-0.219	0.124	0.190	-0.127	-0.120	-0.079	-0.074	-0.018	-0.131	0.171	-0.111	0.014	-0.046
25	TotRoad30	0.068	-0.142	-0.081	-0.077	-0.146	-0.008	-0.004	0.124	0.144	0.145	-0.086	-0.193	0.147	0.171	0.121
26	TotRoadWS	0.146	-0.137	-0.203	-0.341	-0.176	0.076	0.084	0.135	0.226	0.193	-0.098	-0.193	0.148	0.106	0.227

Table I-2 (Continued)

	Variable	Pct_bar	Pct_crop	Pct_for	Pct_past	Pct_urb	Pct_wet	Pct_wtr	PwrPlpTL	Slope3	Slp_mod	Sd_slp	Mean_slp	Soil_er	Strmden	TotRoad30	TotRoadWS
	n =	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86
1	AgPT	-0.130	0.269	-0.414	0.411	0.133	-0.088	0.150	0.059	-0.138	-0.083	-0.178	-0.210	0.098	-0.263	0.068	0.146
2	EPT	-0.127	-0.338	0.488	-0.221	-0.145	-0.250	-0.194	0.001	0.577	0.290	0.697	0.676	0.005	0.292	-0.142	-0.137
3	IBI	0.107	-0.011	-0.023	0.161	-0.263	0.023	0.119	-0.069	-0.125	-0.024	-0.085	-0.108	0.143	-0.219	-0.081	-0.203
4	Hab	0.230	0.118	0.135	-0.291	-0.322	0.210	-0.061	-0.105	-0.131	-0.286	0.192	0.191	-0.459	0.124	-0.077	-0.341
5	Rich	-0.064	-0.180	0.339	-0.172	-0.210	-0.100	-0.164	-0.031	0.268	0.115	0.445	0.417	-0.035	0.190	-0.146	-0.176
1	Ag_hi	0.020	-0.026	-0.031	0.000	0.138	-0.053	-0.072	-0.023	-0.092	-0.270	-0.101	-0.090	0.100	-0.127	-0.008	0.076
2	Ag_slp_hi	0.023	-0.022	-0.041	0.006	0.143	-0.052	-0.069	-0.021	-0.085	-0.261	-0.096	-0.086	0.097	-0.120	-0.004	0.084
3	Ag_mod	-0.451	0.243	-0.586	0.936	0.033	-0.220	0.326	0.148	0.015	0.297	-0.177	-0.221	0.431	-0.079	0.124	0.135
4	Ag_slp	-0.354	0.254	-0.629	0.920	0.660	-0.182	0.298	0.209	0.126	0.275	-0.151	-0.193	0.276	-0.074	0.144	0.226
5	Ag_slp_mod	-0.448	0.131	-0.532	0.928	0.056	-0.210	0.285	0.164	0.187	0.412	-0.100	-0.136	0.414	-0.018	0.145	0.193
6	Bar_slp_hi	0.111	-0.129	0.132	-0.106	-0.022	-0.057	-0.112	-0.041	-0.054	-0.300	-0.097	-0.083	0.101	-0.131	-0.086	-0.098
7	Bar_slp_mod	0.630	-0.223	0.161	-0.253	-0.105	-0.067	0.114	-0.105	0.115	0.197	-0.028	-0.006	0.170	0.171	-0.193	-0.193
8	Crop_slp	-0.071	0.572	-0.670	0.576	0.041	-0.101	0.253	0.395	-0.070	-0.015	-0.252	-0.301	-0.004	-0.111	0.147	0.148
9	Crop_slp_mod	-0.386	0.269	-0.507	0.746	0.066	-0.217	0.256	0.300	0.109	0.370	-0.144	-0.181	0.438	0.014	0.171	0.106
10	Past_slp	-0.430	0.068	-0.513	0.936	0.059	-0.191	0.272	0.091	0.197	0.366	-0.081	-0.114	0.362	-0.046	0.121	0.227
11	Pct_bar	1.000	0.052	0.009	-0.439	-0.159	0.094	0.004	-0.060	-0.252	-0.355	-0.235	-0.226	-0.353	-0.068	-0.151	-0.158
12	Pct_crop	0.052	1.000	-0.735	0.175	-0.063	0.328	0.114	0.228	-0.598	-0.474	-0.416	-0.444	-0.493	-0.296	0.056	0.031
13	Pct_for	0.009	-0.735	1.000	-0.608	-0.319	-0.294	-0.244	-0.199	0.472	0.253	0.492	0.529	0.209	0.323	-0.331	-0.393
14	Pct_past	-0.439	0.175	-0.608	1.000	0.047	-0.187	0.314	0.106	0.044	0.255	-0.162	-0.203	0.369	-0.101	0.118	0.200
15	Pct_urb	-0.159	-0.063	-0.319	0.047	1.000	-0.068	-0.117	0.056	0.049	0.080	-0.061	-0.071	0.116	-0.096	0.541	0.812
16	Pct_wet	0.094	0.328	-0.294	-0.187	-0.068	1.000	0.013	0.003	-0.499	-0.383	-0.267	-0.238	-0.530	-0.154	0.250	-0.089
17	Pct_wtr	0.004	0.114	-0.244	0.314	-0.117	0.013	1.000	-0.107	-0.007	0.091	-0.093	-0.111	0.094	-0.015	-0.104	-0.181
18	PwrPlpTI	-0.060	0.228	-0.199	0.106	0.056	0.003	-0.107	1.000	-0.022	-0.004	-0.053	-0.073	-0.039	0.050	0.203	0.028
19	Slope3	-0.252	-0.598	0.472	0.044	0.049	-0.499	-0.007	-0.022	1.000	0.740	0.764	0.753	0.376	0.487	-0.062	0.026
20	Slp_mod	-0.355	-0.474	0.253	0.255	0.080	-0.383	0.091	-0.004	0.740	1.000	0.368	0.340	0.705	0.462	-0.042	0.074
21	Sd_slp	-0.235	-0.416	0.492	-0.162	-0.061	-0.267	-0.093	-0.053	0.764	0.368	1.000	0.979	-0.013	0.353	0.044	-0.124
22	Mean_slp	-0.226	-0.444	0.529	-0.203	-0.071	-0.238	-0.111	-0.073	0.753	0.340	0.979	1.000	-0.028	0.366	0.054	-0.150
23	Soil_er	-0.353	-0.493	0.209	0.369	0.116	-0.530	0.094	-0.039	0.376	0.705	-0.013	-0.028	1.000	0.201	-0.079	0.044
24	Strmden	-0.068	-0.296	0.323	-0.101	-0.096	-0.154	-0.015	0.050	0.487	0.462	0.353	0.366	0.201	1.000	-0.020	-0.091
25	TotRoad30	-0.151	0.056	-0.331	0.118	0.541	0.250	-0.104	0.203	-0.062	-0.042	0.044	0.054	-0.079	-0.020	1.000	0.566
26	TotRoadWS	-0.158	0.031	-0.393	0.200	0.812	-0.089	-0.181	0.028	0.026	0.074	-0.124	-0.150	0.044	-0.091	0.566	1.000

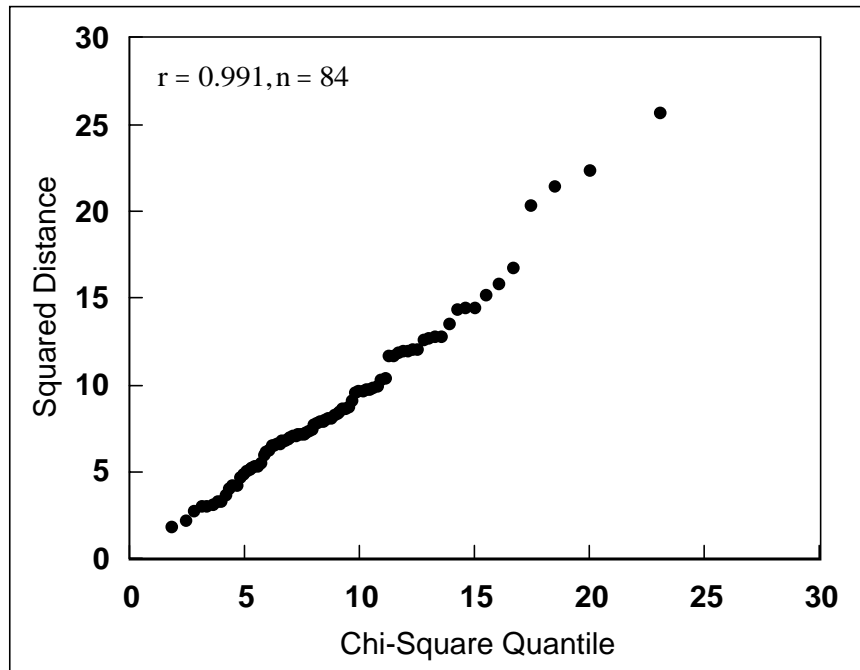


Figure I-1. The squared distance and chi-square quantile for the landscape and biota variables used for the canonical correlation analyses.

3 - Number of Variables to Number of Observations Ratio: The number of variables in each data set must be equal to or more than two (Thorndike, 1978). When the number of variables becomes large, with respect to the number of observations, then it is necessary to consider the ratio of the number of variables to the number of observations in a sample (Gitten, 1980). A high ratio will introduce a bias in estimating the canonical correlation coefficient and roots (eigenvalues). Therefore, the results will be unreliable. A minimum range for the variable to a sample ratio of 0.025-0.05 has been recommended (Barcikowski & Stevens, 1975; Thorndike, 1978). A detailed discussion of the variable/sample ratio is given (Gitten, 1980, Chap. 13). One way to reduce the number of variables is to select the ones that contribute the most to the variability of the data. Therefore, we performed a discriminate analysis with a stepwise selection using the STEPDISC procedure (Proc StepDisc; SAS™) for each of the biological and landscape data sets by ecoregion to reduce the number of variables. SAS™ statements below describe the discriminate procedure.

```
Proc stepdisc data=wl all sle=0.35 sls=0.15;
  class Ecoregion;
  var agpt ibi rich hab ept;
  * var Pct_crop Crop_slp Pct_past Past_slp Pct_wet Pct_bar Pct_wtr
    sd_slp Slope3 Mean_slp TotRoad30 TotRoadWS strmden soil_er;
run;
```

Results for biological parameters were AgPT, IBI, Hab, Rich, and EPT and for landscape variables the results were Past_slp, Pct_bar, Soil_er, and Slope3. The ratio of the number of variables to samples was 0.107 (9/84). Being aware of the external information about the variables and the study area, we concluded that this ratio was satisfactory. The ratios of the number of variables to samples

for water chemistry/landscape metrics and water chemistry/biological metrics were 0.09 (7/77) and 0.009 (7/77), respectively.

Outliers were examined by plotting the x-distance and the y-distance for each site (Appendix I, Figures I-2a and b). There was no evidence of outliers in the data.

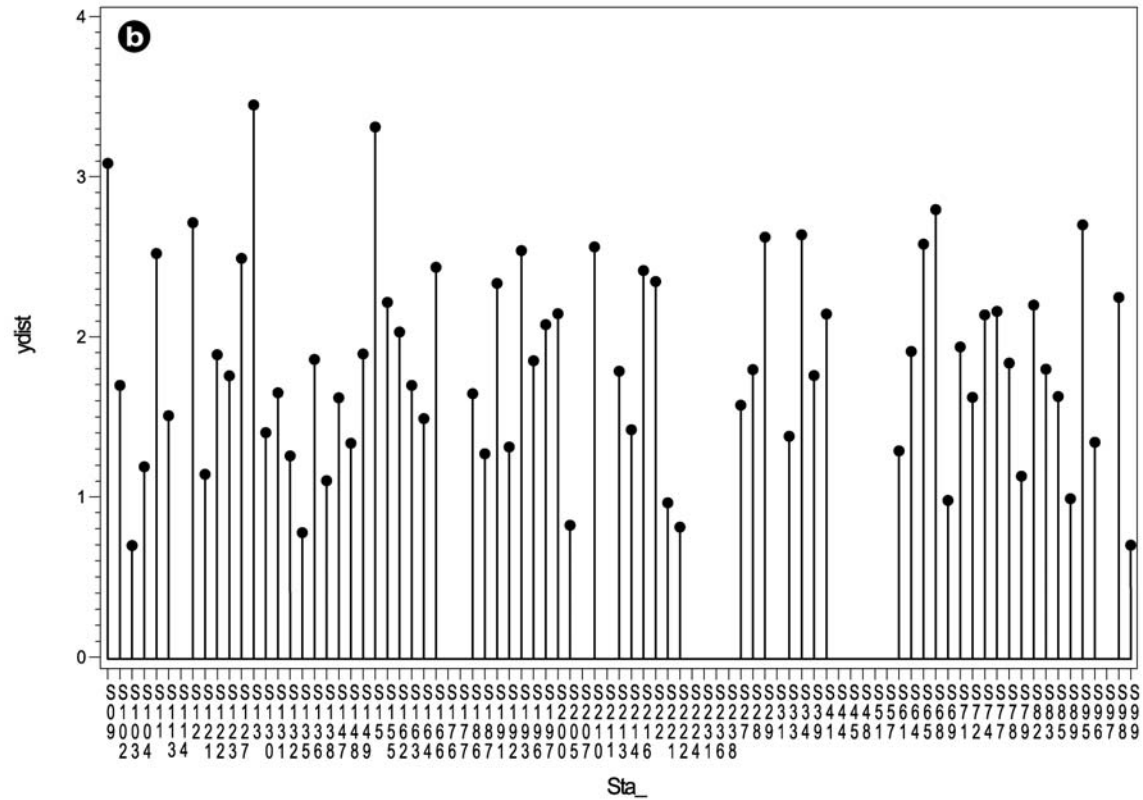
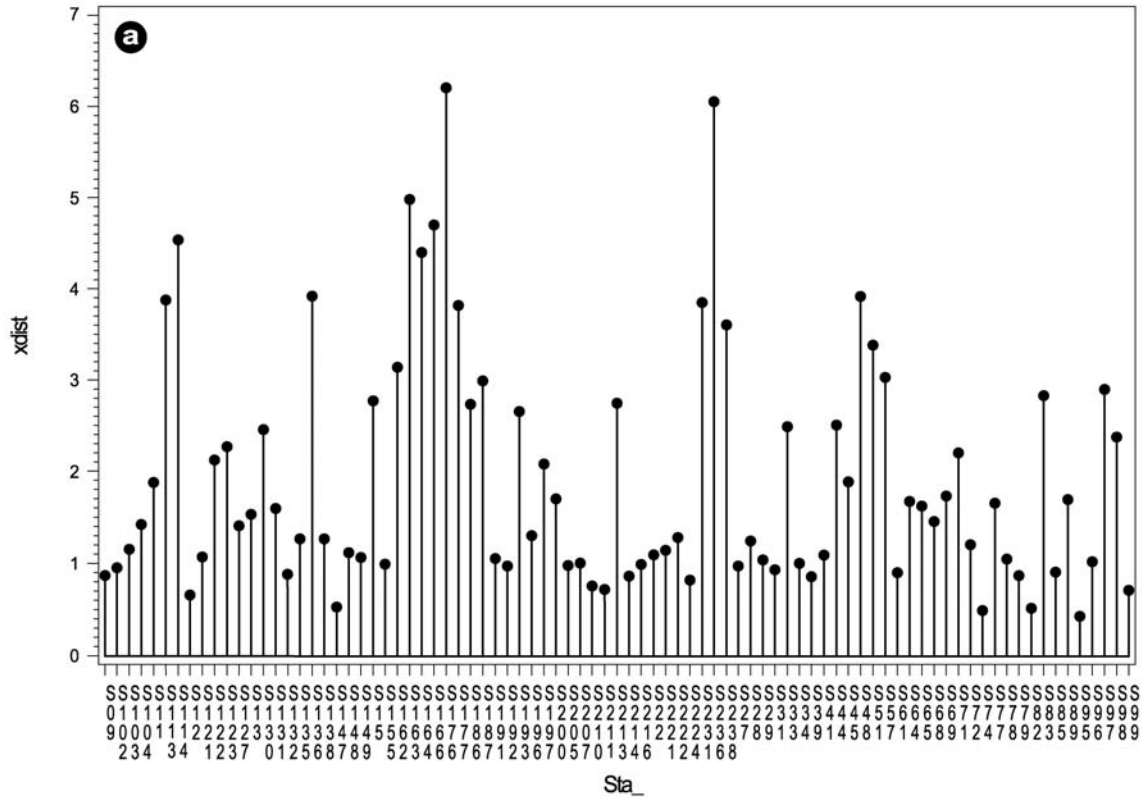


Figure I-2. The x-distance (a) and y-distance (b) to the model.

Appendix II

%multinorm SAS™ Macro

```
**** From SAS *****,
**** Multinormality test, Principal components and factor analyses *****,
%multnorm(data=Chloud.watland,var=AgPT EPT IBI Hab Rich)
*****/

%macro multinorm (
  data=_last_ , /* input data set */
  var= , /* REQUIRED: variables for test */
          /* May NOT be a list e.g. var1-var10 */
  plot=yes , /* create chi-square plot? */
  hires=yes /* high resolution plot? */
);
options nonotes;
%let lastds=&syslast;

/* Verify that VAR= option is specified */
%if &var= %then %do;
  %put ERROR: Specify test variables in the VAR= argument;
  %goto exit;
%end;

/* Parse VAR= list */
%let _i=1;
%do %while (%scan(&var,&_i) ne %str() );
  %let arg&_i=%scan(&var,&_i);
  %let _i=%eval(&_i+1);
%end;
%let nvar=%eval(&_i-1);

/* Remove observations with missing values */
%put NOTE: Removing observations with missing values...;
data _nomiss;
  set &data;
  if nmiss(of &var )=0;
run;

/* Quit if covariance matrix is singular */
%let singular=nonsingular;
%put NOTE: Checking for singularity of covariance matrix...;
proc princomp data=_nomiss outstat=_evals noprint;
```

```

var &var ;
run;
%if &syserr=3000 %then %do;
  %put NOTE: PROC PRINCOMP required for singularity check.;
  %put NOTE: Covariance matrix not checked for singularity.;
  %goto findproc;
%end;
data _null_;
  set _evals;
  where _TYPE_='EIGENVAL';
  if round(min(of &var ),1e-8)<=0 then do;
    put 'ERROR: Covariance matrix is singular.';
    call symput('singular','singular');
  end;
run;
%if &singular=singular %then %goto exit;

%findproc:
/* Is IML or MODEL available for analysis? */
%let mult=yes; %let multtext=%str( and Multivariate);
%put NOTE: Checking for necessary procedures...;
proc iml; quit;
%if &syserr=0 %then %goto iml;
proc model; quit;
%if &syserr=0 and
  (%substr(&sysvlong,1,9)>=6.09.0450 and %substr(&sysvlong,3,2) ne 10)
  %then %goto model;
%put NOTE: SAS/ETS PROC MODEL with NORMAL option or SAS/IML is required;
%put %str( ) to perform tests of multivariate normality. Univariate;
%put %str( ) normality tests will be done.;
%let mult=no; %let multtext=;
%goto univar;

%iml:
proc iml;
  reset;
  use _nomiss; read all var {&var} into _x; /* input data */
  /* compute mahalanobis distances */
  _n=nrow(_x); _p=ncol(_x);
  _c=_x-j(_n,1)*_x[:,,]; /* centered variables */
  _s=(_c*_c)/_n; /* covariance matrix */ _rij=_c*inv(_s)*_c`; /* mahalanobis angles
*/

  /* get values for probability plot and output to data set */
  %if &plot=yes %then %do;
    _d=vecdiag(_rij#(_n-1)/_n); /* squared mahalanobis distances */
    _rank=ranktie(_d); /* ranks of distances */
    _chi=cinv((_rank-.5)/_n,_p); /* chi-square quantiles */
    _chiplot=_d||_chi;
    create _chiplot from _chiplot [colname={'MAHDIST' 'CHISQ'}];
    append from _chiplot;

```

```

%end;

/* Mardia tests based on multivariate skewness and kurtosis */
_b1p=(_rij##3)[+,+]/(_n##2);          /* skewness */
_b2p=trace(_rij##2)/_n;                /* kurtosis */
_k=(_p+1)#(_n+1)#(_n+3)/(_n#((_n+1)#(_p+1)-6)); /* small sample correction */
_b1pchi=_b1p#_n#_k/6;                  /* skewness test statistic */
_b1pdf=_p#(_p+1)#(_p+2)/6;            /* and df */
_b2pnorm=(_b2p-_p#(_p+2))/sqrt(8#_p#(_p+2)/_n); /* kurtosis test statistic */
_prob1p=1-probchi(_b1pchi,_b1pdf);     /* skewness p-value */
_prob2p=2*(1-probnorm(abs(_b2pnorm))); /* kurtosis p-value */

/* output results to data sets */
_names={"Mardia Skewness","Mardia Kurtosis"};
create _names from _names [colname='TEST'];
append from _names;
_probs=(_n||_b1p||_b1pchi||_prob1p) // (_n||_b2p||_b2pnorm||_prob2p);
create _values from _probs [colname={'N' 'VALUE' 'STAT' 'PROB'}];
append from _probs;
quit;

data _mult;
merge _names _values;
run;

%univar:
/* get univariate test results */
proc univariate data=_nomiss noprint;
var &var;
output out=_stat normal=&var ;
output out=_prob probn=&var ;
output out=_n n=&var ;
run;

data _univ;
set _stat _prob _n;
run;

proc transpose data=_univ name=variable
out=_tuniv(rename=(col1=stat col2=prob col3=n));
var &var ;
run;
data _both;
length test $15.;
set _tuniv
%if &mult=yes %then _mult;;
if test=" then if n<=2000 then test='Shapiro-Wilk';
else test='Kolmogorov';
run;
proc print data=_both noobs split='/';
var variable n test %if &mult=yes %then value;

```

```

    stat prob;
title "Univariate&multtext Normality Tests";
label variable="Variable"
    test="Test"    %if &mult=yes %then
value= "Multivariate/Skewness &/Kurtosis";
    stat="Test/Statistic/Value"
    prob="p-value";
run;
%if &plot=yes %then
    %if &mult=yes %then %goto plotstep;
    %else %goto plot;
%else %goto exit;

%model:
/* Multivariate and Univariate tests with MODEL */
proc model data=_nomiss;
    %do _i=1 %to &nvar;
        &&arg&_i = _a;
    %end;
fit &var / normal;
title "Univariate&multtext Normality Tests";
run;
%if &plot ne yes %then %goto exit;

%plot:
/* compute values for chi-square Q-Q plot */
proc princomp data=_nomiss std out=_chiplot noprint;
    var &var ;
run;
%if &syserr=3000 %then %do;
    %put ERROR: PROC PRINCOMP in SAS/STAT needed to do plot.;
    %goto exit;
%end;
data _chiplot;
    set _chiplot;
    mahdist=uss(of prin1-prin&nvar );
    keep mahdist;
run;
proc rank data=_chiplot out=_chiplot;
    var mahdist;
    ranks rdist;
run;
data _chiplot;
    set _chiplot nobs=_n;
    chisq=cinv((rdist-.5)/_n,&nvar);
    keep mahdist chisq;
run;
%plotstep:
/* Create a chi-square Q-Q plot
NOTE: Very large sample size is required for chi-square asymptotics
unless the number of variables is very small.

```

```

*/
%if &hires=yes %then proc gplot data=_chiplot;
%else      proc plot data=_chiplot;;
  plot mahdist*chisq;
  label mahdist="Squared Distance"
        chisq= "Chi-square quantile";
  title "Chi-square Q-Q plot";
  run;
  quit;
%if &syserr=3000 %then %do;
  %put NOTE: PROC PLOT will be used instead.;
  %let hires=no;
  %goto plotstep;
%end;
%exit;
options notes _last_=&lastds;
title;
%mend;
Proc corr data=_chiplot;
  var mahdist chisq;
  run;
  quit;
Proc sort data=_chiplot out=a;
  by mahdist chisq;
  run;
  quit;
Proc plot data = a vpercent=17 hpercent=90;
  plot mahdist*chisq / vref = 12.5 href=13;;
run;
options ps=255 ls=80;
Proc print data = a; /* export this data to other graphing software for better presentation*/
  var mahdist chisq;
run;

```

Appendix III

SAS™ Statements for PLS and Diagnostic Checking

```
libname Factor 'c:\abc' ;

** Principal component to examine strength of Colinearity in the dependent variables ***** ;
Proc Princomp data=chloud.watland out=prin;
  var AgPT EPT IBI Hab Rich;
run;

** Factor analyses to find the number of significant factors for the biota ***** ;
Proc Factor data=chloud.watland method=ML;
  var AgPT EPT IBI Hab Rich;
run;

title1 'Principal components for Bio: AgPT EPT IBI Hab Rich';
title2 'Savannah River Basin ';
%plotit(data=prin, labelvar=sta_,
  Plotvars= prin2 prin1, color=black, colors=blue)
run;
*****;

**** PLS; 26 landscape metrics *****;
Options ps=255 ls=120;
Proc PLS data=Chloud.watland cv=split(9) cvtest(seed=12345);
Model AgPT EPT IBI Hab Rich = Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod Bar_slp_hi
  Bar_slp_mod Crop_slp Crop_slp_mod

  Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet
  Pct_wtr PwrPipTL Slope3 Sd_slp Mean_slp Slp_mod Soil_er Strmden TotRoad30 TotRoadWS;

  Output out=outWLT predicted = yhat1-yhat5
    yresidual = yres1-yres5
    xresidual = xres1-xres26
    xscore = xscr
    yscore = yscr;

run;

***** Diagnostic Check Figures and VIP table *****;
axis1 label=(angle=270 rotate=90 "x score 2")
  major=(number=5) minor=none;
axis2 label=("X-score 1") minor=none;
```

```

Title "Biota(AgPT, EPT, IBI, Hab, Rich) and 26 Landscape Metrics";
symbol1 v=dot c=blue i=none;
proc gplot data=outWLT;
  plot xscr2*xscr1=1
    / vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;
%let ifac = 1;
data pltanno; set outWLT;
  length text $ 5;
  retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
    color 'blue' style 'swissb';
  text=%str(Sta_); x=xscr&ifac; y=yscr&ifac;
  axis1 label=(angle=270 rotate=90 "Y score &ifac")
    major=(number=5) minor=none;
  axis2 label=("X-score &ifac") minor=none;
  symbol1 v=none i=none;
proc gplot data=outWLT;
  plot yscr&ifac*xscr&ifac=1
    / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;

data pltanno; set outWLT;
  length text $ 5;
  retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
    color 'blue' style 'swissb';
  text=%str(Sta_); x=xscr1; y=xscr2;
  axis1 label=(angle=270 rotate=90 "X score 2")
    major=(number=5) minor=none;
  axis2 label=("X-score 1") minor=none;
  symbol1 v=none i=none;
proc gplot data=outWLT;
  plot xscr2*xscr1=1
    / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;

ods output XWeights=xweights;
proc pls data=Chloud.watland nfac=2 details;
Model AgPT EPT IBI Hab Rich = Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod Bar_slp_hi
  Bar_slp_mod Crop_slp Crop_slp_mod

  Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet

  Pct_wtr PwrPipTL Slope3 Sd_slp Mean_slp Slp_mod Soil_er Strmden

  TotRoad30 TotRoadWS;
run;

proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
  out =xweights;
  data xweights; set xweights;

```

```

    rename col1=w1 col2=w2;
data wt_anno; set xweights;
length text $ 7;
retain function 'label'
    position '5'
    hsys '3'
    xsys '2'
    ysys '2'
    color 'blue'
    style 'swissb';
text=%str(_name_); x=w1; y=w2;
run;

axis1 label=(angle=270 rotate=90 "X weight 2")
    major=(number=5) minor=none;
axis2 label=("X-weight 1") minor=none;
symbol1 v=none i=none;
proc gplot data=xweights;
    plot w2*w1=1 / anno=wt_anno vaxis=axis1
        haxis=axis2 frame cframe=white href=0 vref=0;
run; quit;

```

*The following statements produce coefficients and the VIP ;

```

/*
/ Put coefficients, weights, and R**2's into data sets.
/-----*/
ods listing close;
ods output PercentVariation = pctvar
    XWeights = xweights
    CenScaleParms = solution;
proc pls data=Chloud.watland nfac=2 details;
    Model AgPT EPT IBI Hab Rich = Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod
        Bar_slp_hi Bar_slp_mod Crop_slp Crop_slp_mod
        Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet
        Pct_wtr PwrPipTL Slope3 Sd_slp Mean_slp Slp_mod Soil_er Strmden
        TotRoad30 TotRoadWS / solution;
run;
ods listing;

```

```

/*
/ Just reformat the coefficients.
/-----*/
data solution; set solution;
format EPT IBI Hab Rich 8.5;
if (RowName = 'Intercept') then delete;
rename RowName = Predictor AgPT = AgPT;
rename RowName = Predictor EPT = EPT;
rename RowName = Predictor IBI = IBI ;
rename RowName = Predictor Hab = Hab;

```

```

    rename RowName = Predictor Rich= Rich;
run;

/*
/ Transpose weights and R**2's.
/-----*/
data xweights; set xweights; _name_='W'||trim(left(_n_));
data pctvar ; set pctvar ; _name_='R'||trim(left(_n_));
proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
    out =xweights;
proc transpose data=pctvar(keep=_name_ CurrentYVariation)
    out =pctvar;
run;
Proc Print data=pctvar;

/* Updated from previous program
/ Sum the squared weights times the normalized R**2's.
/ The VIP is defined as the square root of this
/ weighted average times the number of predictors.
/-----*/

proc sql;
    create table vip as

    select *,

        w1 /sqrt (uss(w1)) as wnorm1,

        w2 /sqrt (uss(w2)) as wnorm2

    from xweights left join pctvar (drop=_name_) on 1;
data vip; set vip; keep _name_ vip;
array wnorm{2};
array r{2};
VIP = 0;
do i = 1 to 2;

    VIP = VIP + r{1}*(wnorm{i}**2)/sum(of r1-r2);

end;
VIP = sqrt(VIP * 14);
data vipbpls; merge solution vip(drop=_name_);
proc print data=vipbpls;
run;

*Outlier *****;
Proc PLS data=Chloud.watland nfac=1;
Model AgPT EPT IBI Hab Rich = Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod Bar_slp_hi
    Bar_slp_mod Crop_slp Crop_slp_mod

    Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet

```

```

Pct_wtr PwrPipTL Slope3 Sd_slp Mean_slp Slp_mod Soil_er Strmden

TotRoad30 TotRoadWS ;
output out=stdres stdxsse=stdxsse stdysse=stdysse;
data stdres; set stdres;
    xdist = sqrt(stdxsse);
    ydist = sqrt(stdysse);
run;
Symbol1 i=needles v=dot c=blue;
Symbol2 i=needles v=dot c=red;
Symbol3 i=needles v=dot c=green;
Proc gplot data=stdres;
    plot xdist*Sta_=ecoregion / cframe = white ;
Proc gplot data=stdres;
    plot ydist*Sta_=ecoregion / cframe=white;
run;
*****;

**** Refine (Prune) the above model and do the diagnostic checking by graphing ****;
** Use only the VIP with > 0.8 ****;
Proc PLS data=Chloud.watland cv=split(9) cvtest(seed=12345);
    Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
        Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
        Slope3 Sd_slp Mean_slp Soil_er Strmden ;
    Output out=outWLT predicted = yhat1-yhat5
        yresidual = yres1-yres5
        xresidual = xres1-xres14
        xscore = xscr
        yscore = yscr;

run;
axis1 label=(angle=270 rotate=90 "x score 2")
    major=(number=5) minor=none;
axis2 label=("X-score 1") minor=none;
Title "Biota(AgPT, ept, ibi, hab, rich) and 1
4 Landscape Metrics";
symbol1 v=dot c=blue i=none;
proc gplot data=outWLT;
    plot xscr2*xscr1=1
        / vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;
%let ifac = 1;
data pltanno; set outWLT;
    length text $ 5;
    retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
        color 'blue' style 'swissb';
    text=%str(Sta_); x=xscr&ifac; y=yscr&ifac;
axis1 label=(angle=270 rotate=90 "Y score &ifac")
    major=(number=5) minor=none;
axis2 label=("X-score &ifac") minor=none;
symbol1 v=none i=none;

```

```

proc gplot data=outWLT;
  plot yscr&ifac*xscr&ifac=1
  / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;

```

```

data pltanno; set outWLT;
  length text $ 5;
  retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
  color 'blue' style 'swissb';
  text=%str(Sta_); x=xscr1; y=xscr2;
  axis1 label=(angle=270 rotate=90 "X score 2")
  major=(number=5) minor=none;
  axis2 label=("X score 1") minor=none symbol1 v=none i=none;
proc gplot data=outWLT;
  plot xscr2*xscr1=1
  / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;

```

```

ods output XWeights=xweights;
proc pls data=Chloud.watland nfac=2 details;
  Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
  Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
  Slope3 Sd_slp Mean_slp Soil_er Strmden ;
run;

```

```

proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
  out =xweights;
data xweights; set xweights;
  rename col1=w1 col2=w2;
data wt_anno; set xweights;
  length text $ 7;
  retain function 'label'
  position '5'
  hsys '3'
  xsys '2'
  ysys '2'
  color 'blue'
  style 'swissb';
  text=%str(_name_); x=w1; y=w2;
run;

```

```

  axis1 label=(angle=270 rotate=90 "X weight 2")
  major=(number=5) minor=none;
  axis2 label=("X-weight 1") minor=none;
  symbol1 v=none i=none;
proc gplot data=xweights;
  plot w2*w1=1 / anno=wt_anno vaxis=axis1
  haxis=axis2 frame cframe=white href=0 vref=0;
run; quit;

```

*The following statements produce coefficients and the VIP ;

```

/*
/ Put coefficients, weights, and R**2's into data sets.
/-----*/
ods listing close;
ods output PercentVariation = pctvar
      XWeights      = xweights
      CenScaleParms = solution;
proc pls data=Chloud.watland nfac=2 details;
  Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
      Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
      Slope3 Sd_slp Mean_slp Soil_er Strmden / solution;

run;
ods listing;

/*
/ Just reformat the coefficients.
/-----*/
data solution; set solution;
  format EPT IBI Hab Rich 8.5;
  if (RowName = 'Intercept') then delete;
  rename RowName = Predictor AgPT      = AgPT;
  rename RowName = Predictor EPT      = EPT;
  rename RowName = Predictor IBI      = IBI ;
  rename RowName = Predictor Hab      = Hab;
  rename RowName = Predictor Rich     = Rich;
run;

/*
/ Transpose weights and R**2's.
/-----*/
data xweights; set xweights; _name_='W'||trim(left(_n_));
data pctvar ; set pctvar ; _name_='R'||trim(left(_n_));
proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
  out =xweights;
proc transpose data=pctvar(keep=_name_ CurrentYVariation)
  out =pctvar;
run;
Proc Print data=pctvar;

/* Updated from previous program
/ Sum the squared weights times the normalized R**2's.
/ The VIP is defined as the square root of this
/ weighted average times the number of predictors.
/-----*/

proc sql;
  create table vip as

```

```

select *,
    w1 /sqrt(uss(w1)) as wnorm1,
    w2 /sqrt(uss(w2)) as wnorm2

    from xweights left join pctvar (drop=_name_) on 1;
data vip; set vip; keep _name_ vip;
array wnorm{2};
array r{2};
VIP = 0;
do i = 1 to 2;

    VIP = VIP + r{i}*(wnorm{i>**2)/sum(of r1-r2);

end;
VIP = sqrt(VIP * 14);
data vipbpls; merge solution vip(drop=_name_);
proc print data=vipbpls;
run;

*Outlier *****;
Proc PLS data=Chloud.watland nfac=1;
Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 Sd_slp Mean_slp Soil_er Strmden / solution;

    output out=stdres stdxsse=stdxsse stdysse=stdysse;
data stdres; set stdres;
    xdist = sqrt(stdxsse);
    ydist = sqrt(stdysse);
run;
Symbol1 i=needles v=dot c=blue;
Symbol2 i=needles v=dot c=red;
Symbol3 i=needles v=dot c=green;
Proc gplot data=stdres;
    plot xdist*Sta_=ecoregion / cframe = white ;
Proc gplot data=stdres;
    plot ydist*Sta_=ecoregion / cframe=white;
run;
*****;

** Analyses by Ecoregion *****;
*** Data contain missing and non missing biota *****;

Proc PLS data=Chloud.watland cv=split(10) cvtest(seed=12345);
Where Ecoregion='P';
Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 sd_slp mean_slp Soil_er Strmden ;

```

```

Output out=outWLT predicted = yhat1-yhat5
      yresidual = yres1-yres5
      xresidual = xres1-xres14
      xscore   = xscr
      yscore   = yscr;
run;

Proc PLS data=Chloud.watland cv=block(3) cvtest(seed=12345);
  Where Ecoregion='BR';
  Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 sd_slp mean_slp Soil_er Strmden ;
  Output out=outWLT predicted = yhat1-yhat5
    yresidual = yres1-yres5
    xresidual = xres1-xres14
    xscore   = xscr
    yscore   = yscr;
run;

Proc PLS data=Chloud.watland;* cv= cvtest(seed=12345);
  Where Ecoregion='CP';
  Model AgPT EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 sd_slp mean_slp Soil_er Strmden ;
  Output out=outWLT predicted = yhat1-yhat5
    yresidual = yres1-yres5
    xresidual = xres1-xres14
    xscore   = xscr
    yscore   = yscr;
run;

** Can do the diagnostic checking by plotting the figure followed by PLS above *****

```

Appendix IV

SAS™ Statements for Predicting the Nonmeasured Dependent Variables

```

** Create data sets that contain missing and non missing biota for prediction *****;
** 24 Landscape vars *****;

Data Chloud.BioNoMis;
  set Chloud.watland;
  if bio ne 'missing';
  if ecoregion='P';
  keep Sta_ ecoregion AgPT EPT IBI Hab Rich Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod
  Bar_slp_hi Bar_slp_mod Crop_slp Crop_slp_mod
  Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet Pct_wtr PwrPipTL Slope3 sd_slp
  mean_slp Slp_mod Soil_er Strmden TotRoad30 TotRoadWS;
run;
Proc Print data=Chloud.BioNoMis;
  var Sta_ AgPT EPT IBI Hab Rich;
run;
Data Chloud.BioMis;
  set Chloud.watland;
  if bio = 'missing';
  keep Sta_ Ecoregion Ag_hi Ag_slp_hi Ag_mod Ag_slp Ag_slp_mod
  Bar_slp_hi Bar_slp_mod Crop_slp Crop_slp_mod
  Past_slp Pct_bar Pct_crop Pct_for Pct_past Pct_urb Pct_wet
  Pct_wtr PwrPipTL Slope3 sd_slp mean_slp Slp_mod Soil_er Strmden
  TotRoad30 TotRoadWS;
run;
Proc Print data=Chloud.BioMis;
  Var Sta_ Ecoregion;
run;
Data Chloud.Bio;
  set Chloud.watland;
  ObsAgPT= AgPT;
  Obsept = EPT;
  Obsibi = IBI;
  ObsHab = Hab;
  ObsRich= Rich;
  if bio = 'missing';
  keep Ecoregion Sta_ ObsAgPT Obsept Obsibi ObsHab ObsRich;
run;
Proc Print data=Chloud.Bio;
run;

```

```

***** Prediction for the Piedmont only; considering VIP *****;
** Data without missing bio data, no AgPT, sd_slp and mean_slp ; Oct 10 2002; nLS=12 **;

Proc PLS data=Chloud.BioNoMis cv=Block(14) cvtest(seed=12345); ** split(15) Block(14,30,45) **;
  Model EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
        Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
        Slope3 Soil_er Strmden;
  Output out=outWLT predicted = yhat1-yhat4

      yresidual = yres1-yres4

      xresidual = xres1-xres12
                xscore  = xscr

                yscore  = yscr;

run;
Proc Print data=outWLT;
  Var Sta_ Ecoregion EPT yhat1 IBI yhat2 Hab yhat3 Rich Yhat4;
run;
Symbol1 v=dot i=none c=blue;
Proc Gplot data=outWLT;
  Plot EPT*Yhat1 =1;
  Plot IBI*Yhat2 =1;
  Plot Hab*Yhat3=1;
  Plot Rich*Yhat4=1;
run;

%let ifac = 1;
data pltanno; set outWLT;
  length text $ 5;
  retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
        color 'blue' style 'swissb';
  text=%str(Sta_); x=xscr&ifac; y=yscr&ifac;
  axis1 label=(angle=270 rotate=90 "Y score &ifac")
        major=(number=5) minor=none;
  axis2 label=("X-score &ifac") minor=none;
  symbol1 v=none i=none;
proc gplot data=outWLT;
  plot yscr&ifac*xscr&ifac=1
        / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;
data pltanno; set outWLT;
  length text $ 5;
  retain function 'label' position '5' hsys '3' xsys '2' ysys '2'
        color 'blue' style 'swissb';
  text=%str(Sta_); x=xscr1; y=xscr2;
  axis1 label=(angle=270 rotate=90 "X score 2")
        major=(number=5) minor=none;
  axis2 label=("X-score 1") minor=none;
  symbol1 v=none i=none;

```

```

proc gplot data=outWLT;
  plot xscr2*xscr1=1
    / anno=pltanno vaxis=axis1 haxis=axis2 frame cframe=white href=0 vref=0;
run;
ods output XWeights=xweights;
proc pls data=Chloud.BioNoMis nfac=2 details;
  Model EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 Soil_er Strmden;
run;
Proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
  out =xweights;
data xweights; set xweights;
  rename col1=w1 col2=w2;
data wt_anno; set xweights;
  length text $ 7;
  retain function 'label'
    position '5'
    hsys '3'
    xsys '2'
    ysys '2'
    color 'blue'
    style 'swissb';
  text=%str(_name_); x=w1; y=w2;
run;

axis1 label=(angle=270 rotate=90 "X weight 2")
  major=(number=5) minor=none;
axis2 label=("X-weight 1") minor=none;
symbol1 v=none i=none;
proc gplot data=xweights;
  plot w2*w1=1 / anno=wt_anno vaxis=axis1
    haxis=axis2 frame cframe=white href=0 vref=0;
run; quit;

***** The following statements produce coefficients and the VIP ;
/*
/ Put coefficients, weights, and R**2's into data sets.
/-----*/
ods listing close;
ods output PercentVariation = pctvar
  XWeights = xweights
  CenScaleParms = solution;
proc pls data=Chloud.BioNomis nfac=8 details;
  Model EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 Soil_er Strmden / solution;
run;
ods listing;

/*

```

```

/ Just reformat the coefficients.
/-----*/
data solution; set solution;
  format EPT IBI Hab Rich 8.5;
  if (RowName = 'Intercept') then delete;
  rename RowName = Predictor EPT = EPT;
  rename RowName = Predictor IBI = IBI ;
  rename RowName = Predictor Hab = Hab;
  rename RowName = Predictor Rich= Rich;
run;

/*
/ Transpose weights and R**2's.
/-----*/
data xweights; set xweights; _name_='W'||trim(left(_n_));
data pctvar ; set pctvar ; _name_='R'||trim(left(_n_));
proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
  out =xweights;
proc transpose data=pctvar(keep=_name_ CurrentYVariation)
  out =pctvar;
run;

/* Updated from previous program
/ Sum the normalized squared weights times the normalized R**2's.
/ The VIP is defined as the square root of this
/ weighted average times the number of predictors.
/-----*/
options ps=255 ls=120;
Title ' VIP for the piedmont, nonmissing: nSite=52, nLS=12';
proc sql;

  create table vip as

  select *,

  w1 / sqrt(uss(w1)) as wnorm1,

  w2 / sqrt(uss(w2)) as wnorm2,

  w3 / sqrt(uss(w3)) as wnorm3,

  w4 / sqrt(uss(w4)) as wnorm4,

  w5 / sqrt(uss(w5)) as wnorm5,

  w6 / sqrt(uss(w6)) as wnorm6,

  w7 / sqrt(uss(w7)) as wnorm7,

  w8 / sqrt(uss(w8)) as wnorm8

```

```

        from xweights left join pctvar (drop=_name_) on 1;
data vip; set vip; keep_name_ vip;
  array wnorm {8};
  array r {8};
  VIP = 0;
  do i = 1 to 8;

    VIP = VIP + r{i}*(wnorm{i}**2)/sum(of r1-r8);

  end;
  VIP = sqrt(VIP * 12);
data vipbpls; merge solution vip(drop+_name_);
proc print data=vipbpls;
run;

*Outlier *****;
Proc PLS data=Chloud.BioNoMis nfac=1;
  Model EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 Soil_er Strmden;
  output out=stdres stdxsse=stdxsse stdysse=stdysse;
  data stdres; set stdres;
    xdist = sqrt(stdxsse);
    ydist = sqrt(stdysse);
run;
Symbol1 i=needles v=dot c=blue;
Proc gplot data=stdres;
  plot xdist*Sta_=1 / cframe = white ;
Proc gplot data=stdres;
  plot ydist*Sta_=1 / cframe=white;
run;
*****;

**** Prediction of the missing bio data Oct 2002, nLS = 12 ****;
Title ' Predicted for missing observations, Piedmont ';
Options ps=255 ls=180;
Data Chloud.all; Set Chloud.BioNoMis Chloud.BioMis;
run;
Proc PLS data=Chloud.all cv=block(14) cvtest(seed=12345);
  Model EPT IBI Hab Rich = Ag_mod Ag_slp Ag_slp_mod
    Crop_slp Crop_slp_mod Past_slp Pct_bar Pct_for Pct_past
    Slope3 Soil_er Strmden;
  Output out=Pred p=Predept Predibi PredHab PredRich;
run;
Proc Print data=pred;
  Where (Sta_ in ('S04','S114','S176','S177','S207','S211','S224','S231',
    'S236','S238','S31','S44','S45','S48','S51','S57','S89','S97'));
  var Ecoregion Sta_ PredEPT PredIBI PredHab PredRich;
run;
***** End of prediction *****;

```




United States
Environmental Protection
Agency

Office of Research
and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
\$300

EPA/600/X-05/004
August 2005
www.epa.gov

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE ; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED
STANDARD
POSTAGE
AND FEES
PAID



Recycled/Recyclable
Printed with vegetable-based ink on
paper that contains a minimum of
50% post-consumer fiber content
processed chlorine free