

Partial Least Square Analyses for Association of Landscape Metrics with Surface Water Biological and Chemical Properties in the Savannah River Basin



Maliha S. Nash and Deborah J. Chaloud

14th EPA Conference on Statistics and Information, May 14-17, 2001, Philadelphia, PA

Landscape Ecology Branch, U.S. EPA, Office of Research and Development, National Exposure Research Laboratory

Environmental Sciences Division, P.O. Box 93478, Las Vegas, NV 89193-3478

1. INTRODUCTION

Surface water quality is related to conditions in the surrounding geophysical environment, including soils, landcover, and anthropogenic activities. A number of statistical methods may be used to analyze and explore relationships among variables. Single-, multiple- and multivariate regression analyses have been used to relate water nutrient concentrations to selected landscape metrics. Partial Least Square (PLS) is a multivariate analysis used to explore relationships between two data sets and predict variability for each data set. PLS can also predict values of (or estimates of) the dependent variables that are not sampled in new locations or where the measured independent variables may be highly correlated.

2. DATA USED

The water data used in this analysis were 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 the purpose of this poster, we only used water biology (Biota) from stream reaches centered around the point sites, and landscape metrics (LS) generated for the drainage areas to the point sites. The analyses by the ecoregion were of particular interest and that is what is presented here. For each of the selected sites, the watershed support area was delineated and a suite of landscape metrics were calculated. Specific variables used are shown in Table 1.

4. RESULTS

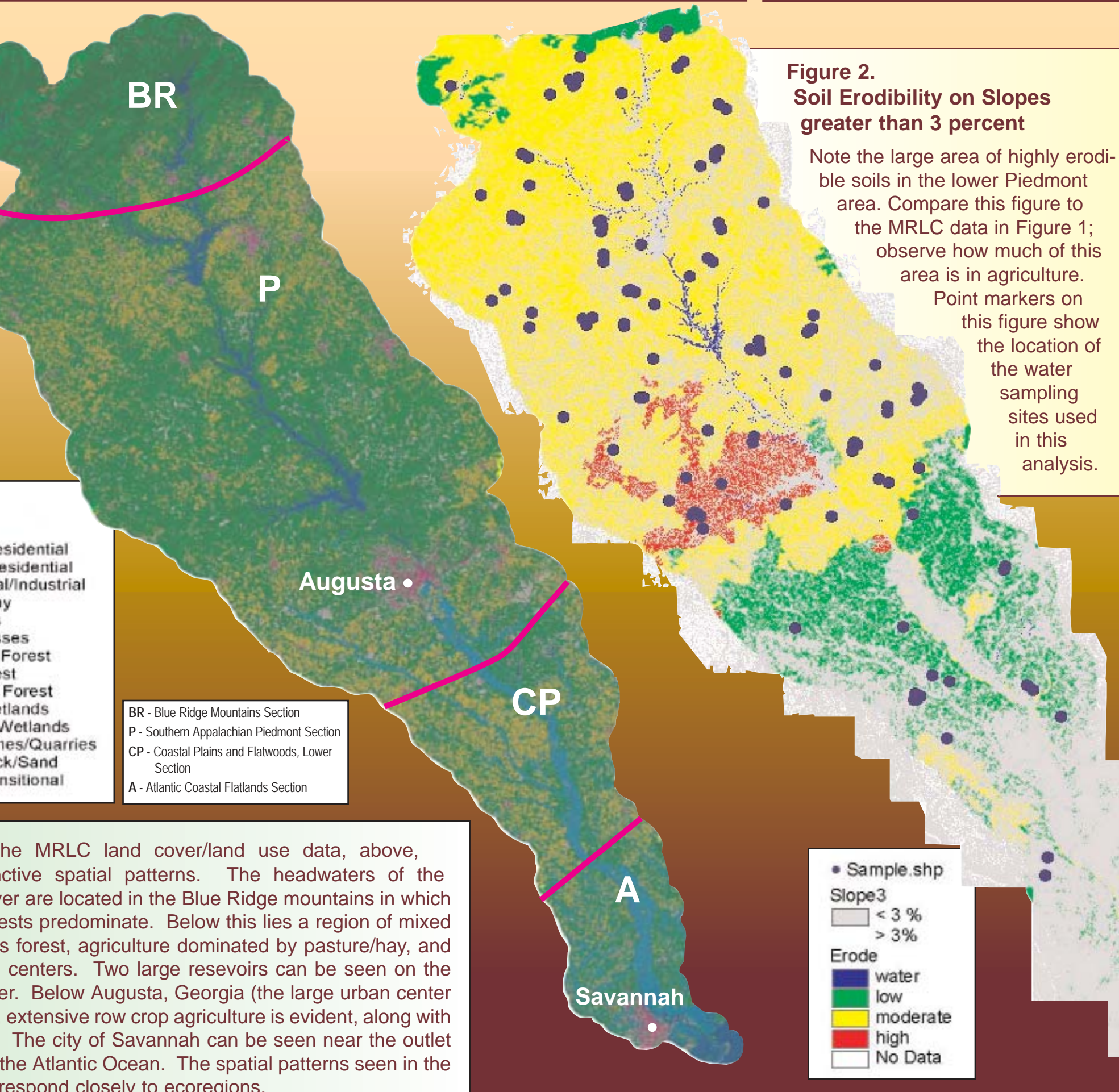


Figure 1. The MRLC land cover/land use data, above, reveals distinctive spatial patterns. The headwaters of the Savannah River are located in the Blue Ridge mountains in which evergreen forests predominate. Below this lies a region of mixed and deciduous forest, agriculture dominated by pasture/hay, 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 wetland area. The city of Savannah can be seen near the outlet of the river to the Atlantic Ocean. The spatial patterns seen in the landcover correspond closely to ecoregions.

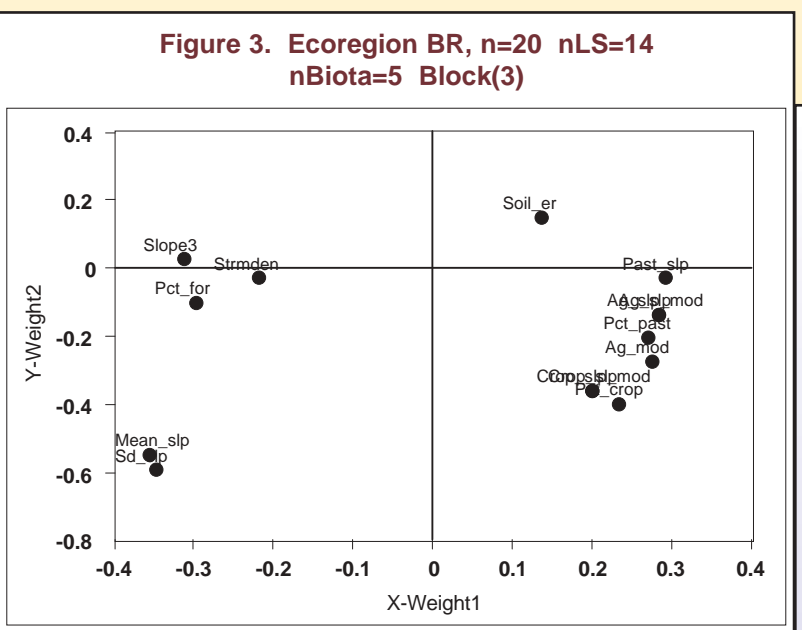
Figure 2. Soil Erodibility on Slopes greater than 3 percent. Note the large area of highly erodible soils in the lower Piedmont area. Compare this figure to the MRLC data in Figure 1; observe how much of this area is in agriculture. Point markers on this figure show the location of the water sampling sites used in this analysis.

Table 1. Variable Descriptions.

Biological Variables	
Biological Integrity	
Mi_ept	Microinvertebrate (Mi), Stream Insect (ept)
Mi_hab	Microinvertebrate (Mi), habitat
Mi_rich	Microinvertebrate (Mi), richness
Fish_ibi	fish index of biological integrity.
Trophic Condition	
AgPT	Algal percent, measure of nutrient enrichment
Landscape Metrics	
Ag_mod	Agriculture on moderately erodible soils
Ag_slp	Agriculture on slopes >3%
Ag_slp_mod	Agriculture on slopes >3% on moderately erodible soils
Crop_slp	Crop on slopes >3%
Crop_slp_mod	Crop on slopes >3% on moderately erodible soils
Past_slp	Pasture on slopes >3%
Pct_for	Percent forest
Pct_crop	Percent crop
Pct_past	Percent pasture
Slope3	Slope >3%
Soil_er	Erodible soils
strmden	Stream density
sd_slp	Standard Deviation of percent slope
Mean_slp	Mean percent slope

3. HOW PLS PERFORMS

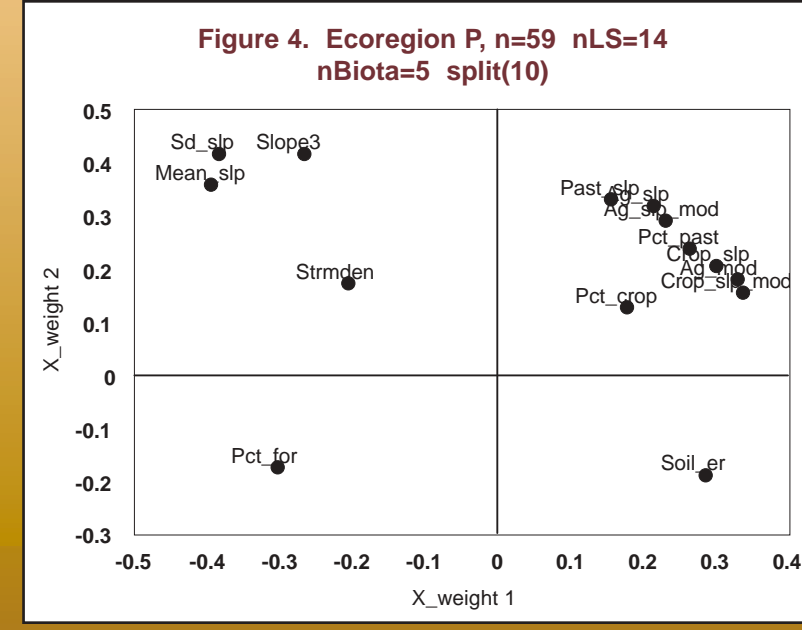
In Partial Least Square of Projection (PLS) analysis, both data sets were first centered and scaled (e.g., Bio0 and Land0). Then a linear combination was composed on the independent variables ($v = \text{Bio}0$; w is the score and w is weight) and the dependent variables ($u = \text{Land}0$; u is the score and t is weight). The linear composition on each data sets is build in to maximize the **covariance** between them. These linear combinations are called factors. PLS extracts many factors from the data sets; the first factor is explained above, the second uses the residuals from the first factor to find the linear combinations of both data sets such that their covariance is maximized. The process is repeated by taking residuals from the previous factor, producing $n-1$ factors ($n = \text{total number of observations}$).



a) Blue Ridge "BR"
There was only one significant factor which accounted for 17% and 74% of the variability for the biota and landscape metrics, respectively. Ag_mod, Ag_slp, Ag_slp_mod, Past_slp, percent forest, percent pasture, total area with slope >3%, sd_slp, Mean_slp (VIP >1; Table 2 & Figure 3) were the most important variables followed by the percent crop and stream density ($0.8 < \text{VIP} < 1.0$; Table 2 & Figure 3). Erodeable soil, crop on slopes >3, and crops on areas with slope >3% and on moderately erodible soil were less important (VIP <0.8; Table 10; Table 2 & Figure 3). The Blue Ridge 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. Stream density in the Blue Ridge is the greatest of the three ecoregions comprising the Savannah Basin. 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% of the land area.

Table 2. Coefficients of the biota and Variable Influence on Projection (VIP) for landscape metrics for Blue Ridge "BR" ecoregion. Red indicates the highest value. Blue and Green indicate the next highest values.

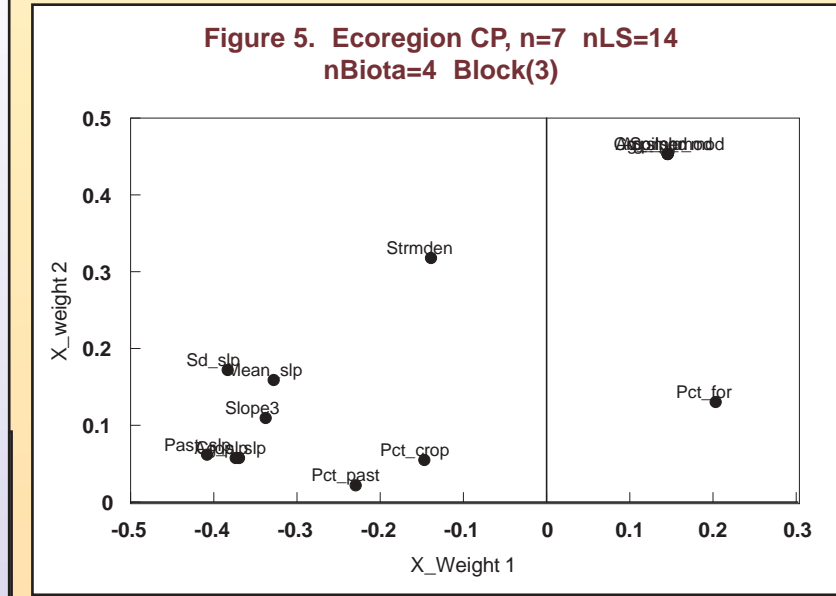
Predictor	AgPT	Mi_ept	Fish_ibi	Mi_hab	Mi_rich	VIP
Ag_mod	0.04174	-0.05052	0.01728	-0.01603	-0.03799	1.023
Ag_slp	0.04330	-0.05240	0.01793	-0.01663	-0.03940	1.062
Ag_slp_mod	0.04332	-0.05242	0.01793	-0.01664	-0.03942	1.062
Crop_slp	0.03052	-0.03694	0.01264	-0.01172	-0.02778	0.748
Crop_slp_mod	0.043061	-0.03705	0.01267	-0.01172	-0.02786	0.750
Past_slp	0.04445	-0.05381	0.01841	-0.01708	-0.04046	1.090
Pct_for	-0.04547	0.05504	-0.01883	0.01747	0.04136	1.115
Pct_crop	0.03556	-0.04303	0.01472	-0.01366	-0.03236	0.872
Pct_past	0.04115	-0.04981	0.01704	-0.01581	-0.03745	1.009
Slope3	-0.04780	0.05785	-0.01979	0.01836	0.04350	1.172
Soil_er	0.02091	-0.02531	0.00866	-0.00803	-0.01903	0.513
strmden	-0.03369	0.04078	-0.01395	0.01294	0.03066	0.826
sd_slp	-0.05318	0.06437	-0.02202	0.02043	0.04840	1.304
Mean_slp	-0.05428	0.06570	-0.02247	0.02085	0.04940	1.331



b) Piedmont "P"
There were three significant factors that accounted for 29% and 86% of the variability for the biota and landscape metrics, respectively (Table 3). The slopes, percent of the total area on slope >3%, and erodible soil (VIP >1) were the most important variables (Table 3). 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 4 ($0.8 < \text{VIP} < 1.0$); they have positive impact on factor 1 and on factor 2. Stream density is less important here than in the Blue Ridge ecoregion. Over half of the Piedmont ecoregion is terrain with slopes greater than 3%. 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% is evident throughout the ecoregion. All of the highly erodible soils in the basin are in this ecoregion; only very small patches of low-erodible soils occur in the Piedmont, generally along the outer edge of the basin. Percent forest and Stream density is generally less than that of the Blue Ridge, but much greater than that of the Coastal Plains. Agriculture land uses are correlated positively with the biology 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 3. Coefficients of the biota and Variable Influence on Projection (VIP) for landscape metrics for Piedmont "P" ecoregion. Red indicates the highest value. Blue and Green indicate the next highest values.

Predictor	AgPT	Mi_ept	Fish_ibi	Mi_hab	Mi_rich	VIP
Ag_mod	0.06581	-0.05308	0.00437	-0.04933	-0.04007	0.994
Ag_slp	0.05415	0.02204	-0.02300	-0.03363	-0.00086	0.892
Ag_slp_mod	0.0433	0.00505	-0.01496	-0.05919	-0.02000	0.899
Crop_slp	0.06995	-0.03400	-0.00337	-0.03011	-0.02315	0.943
Crop_slp_mod	0.04234	-0.07070	0.01432	-0.09762	-0.07002	1.035
Past_slp	0.04144	0.04328	-0.02874	-0.03137	0.00866	0.815
Pct_for	-0.09665	0.03513	0.00599	-0.02514	-0.00167	0.991
Pct_crop	0.12532	0.00979	-0.02438	0.14737	0.07553	1.084
Pct_past	0.06248	-0.01445	-0.00976	-0.03102	-0.010537	0.885
Slope3	-0.05961	0.18879	-0.06605	-0.03760	0.06369	1.261
Soil_er	-0.04989	-0.17296	0.07080	-0.21475	-0.17145	1.226
strmden	-0.02735	0.11329	-0.03899	0.02535	0.05918	0.683
sd_slp	-0.0691	0.24746	-0.09222	0.10386	0.15279	1.373
Mean_slp	-0.0279	0.23252	-0.08418	0.09314	0.14086	1.325



c) Coastal Plain "CP"
The scarcity of sampling sites ($n=7$) in the Coastal Plain precludes an analysis like those completed for the Blue Ridge and Piedmont areas (Table 4). However, erodible soil and all agriculture/soil/slope-related landscape metrics yielded VIPs greater than 1 (Table 4; Figure 5). Percent forest, percent crop, and percent pasture were the least important (VIP <0.8; Table 4) landscape metrics in the Coastal Plain. Soils in this ecoregion are generally of low erodibility, and the terrain is much flatter than the other two ecoregions, hence, area on slope >3% was not significant as in the Blue Ridge and Piedmont ecoregions. Much of the agriculture is in row crops which are subject to run off, particularly when located on slopes and/or erodible soils. So, while agriculture on slopes and/or moderately erodible soils represents a relatively small percentage of the total landcover, these metrics may cause significant impacts on stream biology on a local scale. Although not significant, all landscape metrics correlated positively with AgPT (Table 4) suggesting, as in the Piedmont, these landscape metrics may be indicative of sources of nutrient inputs to streams.

Table 4. Coefficients of the biota and Variable Influence on Projection (VIP) for landscape metrics for Coastal Plain "CP" ecoregion. Red indicates the highest value. Blue and Green indicate the next highest values.

Predictor	AgPT	Mi_ept	Mi_hab	Mi_rich	VIP
Ag_mod	0.200	-0.026	0.085	0.015	1.188
Ag_slp	0.012	0.125	-0.105	0.090	1.064
Ag_slp_mod	0.200	-0.026	0.085	0.015	1.188
Crop_slp	0.012	0.124	-0.104	0.089	1.053
Crop_slp_mod	0.200	-0.026	0.085	0.015	1.188
Past_slp	0.014	0.137	-0.115	0.099	1.164
Pct_for	0.065	-0.059	0.072	-0.031	0.654
Pct_crop	0.018	0.051	-0.039	0.039	0.442
Pct_past	0.001	0.076	-0.066	0.054	0.651
Slope3	0.035	0.116	-0.090	0.088	0.991
Soil_er	0.200	-0.026	0.085	0.015	1.188
strmden	0.131	0.061	-0.012	0.065	0.882
sd_slp	0.060	0.134	-0.097	0.104	1.165
Mean_slp	0.056	0.115	-0.082	0.090	1.006

5. CONCLUSION

In spite of small sample size, PLS permitted analyses of the data by ecoregion, an option is not available with other multivariate analyses (e.g., canonical correlation). The analyses revealed that different landscape metrics affect surface water biota based on their spatial association (e.g., Ecoregion). Percent forest, percent total area on slope with >3%, and slopes are the most important landscape variables in Blue Ridge; percent of total area on slope >3%, soil erodibility, percent crop, and Crop_slp_mod were the most important in Piedmont; soil erodibility, Ag_slp_mod and pasture on slopes were the most important landscape variables in Coastal Plain. Stream density was more important in the Blue Ridge than in the Piedmont. Erodeable soil was the common landscape variable in Piedmont and Coastal Plain. Model Performance was best at the Piedmont ecoregion and this model may be used to predict the biota in other locations from landscape metrics.

Note: Detailed description of PLS statistical analysis and its application are given in EPA Technical Report, EPA/600/R-02/091, November 2002.