Using Canonical Correlation to Detect Associations of Landscape Metrics with Water Biological and Chemical Properties in the Savannah River Basin

Deborah J. Chaloud and Maliha S. Nash

Above and Beyond 2001; EPA Remote Sensing Conference, 20th - 21st, March 2001, Las Vegas, NV

U.S. EPA, Office of Research and Development, National Exposure Research Laboratory, Environmental Sciences Division, P.O. Box 93478, Las Vegas, NV 89193-3478

Surface water quality is related to conditions in the surrounding geophysical environment, including soils, landcover, and anthropogenic activities. Man's activities, from dam building and road construction to growth of cities and industrial pollution, can profoundly impact surface water chemical, physical, and biological qualities.

A number of statistical methods may be used to analyze and explore relationships among variables. Single- and multiple-regression analysis has frequently been used to relate water nutrient concentrations to selected landscape metrics. Canonical correlation, however, is better suited to exploring the relationships among two or more distinct data sets to describe their association and connection to the physical environment.

DATA SETS

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 each of the selected sites, the watershed support area was delineated (see Figure 2) and a suite of landscape metrics were calculated. Specific variables used in the canonical correlations were:

Water chemistry: pH, conductivity (EC), and dissolved oxygen (DO), Water biology: Microinvertebrate (Mi) habitat, Mi_ept, Mi_richness, fish index of biological integrity, and Algal growth Potential Test (AgPT) and Landscape metrics: Percent (Pct)_bar, Pct_for, Past_slp, Soil_er and Slope > 3 percent.

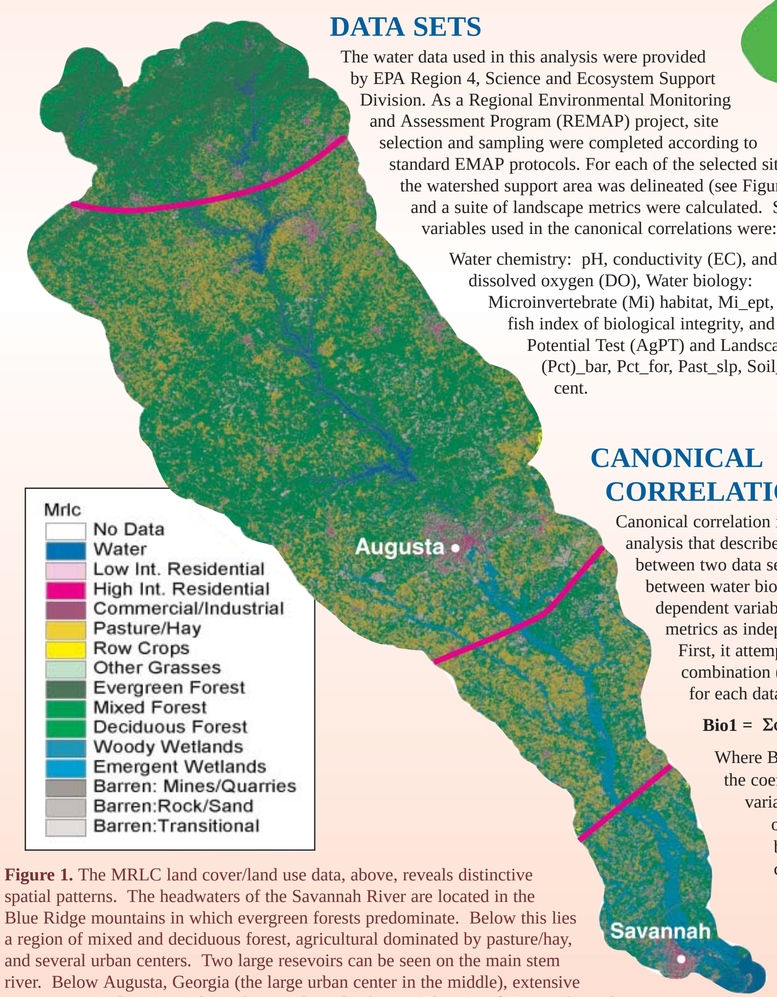

Figure 1. The MRLC land cover/land use data, above, reveals distinctive spatial patterns. The headwaters of the Savannah River are located in the Blue Ridge mountains in which evergreen forests predominate. Below this lies a region of mixed and deciduous forest, agricultural dominated by pasture/hay, and several urban centers. Two large resevoirs 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, as shown in Figure 2.

CANONICAL CORRELATION

Canonical correlation is a multivariate analysis that describes the association between two data sets (e.g., the relation between water biological metrics as dependent variables and landscape metrics as independent variables). First, it attempts to derive a linear combination (*canonical variate*) for each data set as:

$$\textbf{Bio1} = \Sigma\varphi_i B_i \qquad \textbf{Land1} = \Sigma\omega_i LS_i$$

Where Bio1 and Land1 are the first pair of biological and landscape variates; φ_i and ω_i are the coefficients (sometimes refereed as weights) for each of the biological and landscape variables in the first pair of canonical variate; B and LS are the biological and landscape original variables. Values of the coefficients, φ_i and ω_i, are determined so the correlation between Bio1 and Land1 is the maximum. The first pair of variate always has the highest canonical correlation (r_k).

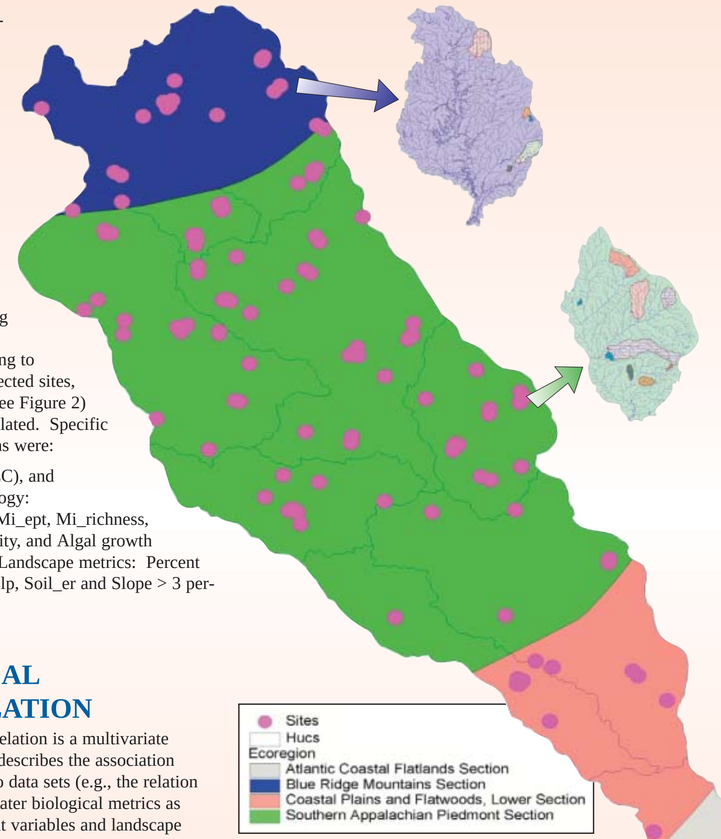

Figure 2. Sample locations for the study area.

Note: Detailed description about variable selection, colinearity, multinormality and other statistical issues for canonical correlation analyses are given in EPA technical report, EPA/600/R-02/091, November 2002.

RESULTS

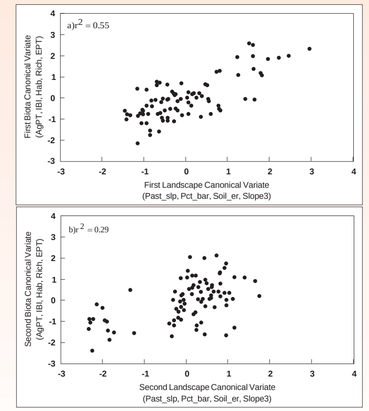

Figure 3. The fitted model consists of the first two canonical variates ($p \leq 0.0008$). The strength of the linear relationship (canonical correlation; r) between the first and second canonical variates of the landscape metrics and surface water biological variables was 0.74 and 0.53, respectively. The first pair of canonical variates explained 0.55 (r^2) of the variation in their linear combinations. Total amount of variance that shared and predicted by the two data sets was 0.92. The fitted model, which contains only the first two canonical variates, accounted for 90% of that shared variance.

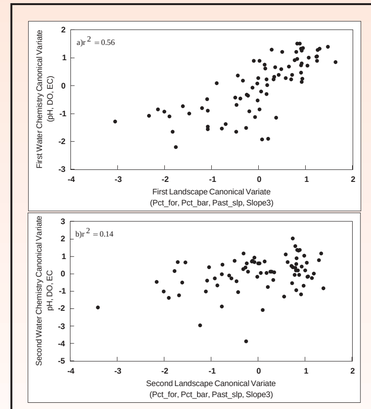

Figure 4. The fitted model consists of the first two canonical variates ($p \leq 0.0134$). The strength of the linear relationship (canonical correlation; r) between the first and second canonical variates of the landscape metrics and surface water chemistry variables was 0.75 and 0.38, respectively. The first pair of canonical variates explained 0.56 (r^2) of the variation in their linear combinations. Total amount of variance that shared and predicted by the two data sets was 0.77 (total r^2). The fitted model, which contains only the first two canonical variates, accounted for 91% of that shared variance.

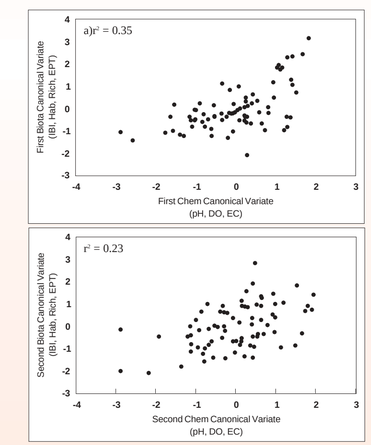

Figure 5. The fitted model consists of Bio1-Chem1 and Bio2-Chem2 ($p \leq 0.0034$). The strength of correlation between Bio1 and Chem1 was 0.54, and between Bio2 and Chem2 was 0.48. The amount of variation in Bio1 attributable to Chem1 was 60% and the amount of variation in Bio2 attributable to Chem2 was 39%. Variance shared by all three canonical variates was 0.59. Our fitted model accounted for 99% of the variance that was shared by all canonical variates.

The landscape-biota model indicated three major contributing variables: the landscape metrics total areas with slope greater than 3 percent (Slope3), the Bio variable EPT (an indicator of three microinvertebrate genera), and the Bio variable Rich (an index of microinvertebrate species richness). Area with the slopes greater than 3 percent was also the major contributing landscape metric in the landscape-chemistry model; the major 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 chemistry variables, with EC negatively related to biota.

The strength of landscape and chemistry relationships was higher than that of landscape and biota. In circumstances where biota data are unavailable or cannot be measured because of costs or other restrictions, the condition of the surface water biological properties can be assessed using chemistry data as surrogates.

CONCLUSION

Canonical correlation analyses indicated increased slope (indicating complex topography, generally occurring in the mountainous areas of the Savannah River Basin) is associated with increased microinvertebrate quality and higher DO concentrations, while the percentage of landcover in pasture on slopes greater than 3 percent is associated with increased conductivity and declines in aquatic biota quality.