Contribution of Nutrients and *E. coli* to Surface Water Condition in the Ozarks

Part I. Using Partial Least Squares Predictions When Standard Regression Assumptions Are Violated

Maliha S. Nash* and Ricardo D. Lopez

U.S. EPA, 944 East Harmon Avenue, Las Vegas, Nevada 89119

**Corresponding author e-mail: nash.maliha@epa.gov*

1. Introduction

Investigation of associations among constituents of surface water and landscapes involves statistical analyses of fundamentally different data sets. Data on surface water conditions are generally obtained through field sampling programs and field/analysis programs are expensive and labor intensive; consequently, the total number of sample sites is usually small. The data set may contain missing values due to the realities of sampling or cost. Landscape data, however, is derived from remote sensing platforms, thereby permitting wall-to-wall coverage. The landscape data sets may contain a very large number of variables, although many of these are not wholly independent (i.e., they may be collinear). Single- and multiple-regression analysis has frequently been used to relate water nutrient concentrations to selected landscape variables that are sensitive to missing values and dependence of predictors (landscape variables). Reliable statistically significant results generally cannot be obtained unless the total number of samples greatly exceeds the number of variables. Partial least squares (PLS) analysis offers a number of advantages over the more traditionally used regression analyses. It has been found to be useful both for providing accurate predictions and for interpreting relationships between data sets containing a high degree of collinearity (see references in Nash et al., 2005).

2. Data Description

2.1. SITE DESCRIPTION

The study area is in the Upper White River study area (21,848 km²) in the Ozarks of Missouri and Arkansas (Figure 1).

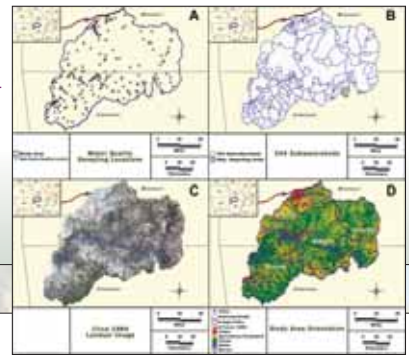

Figure 1. The Upper White River study area is in the Ozarks of Missouri and Arkansas, where 244 water quality sampling locations were sampled (A) and used as "pour points," from which 244 contributing subwatersheds were delineated (B). A combination of multiple Landsat Thematic Mapper imagery (C) and digital aerial photography was used to produce a 2000 land cover map of the study area (D), which was used to calculate landscape metrics.

2.2. DATA

For each of the selected sites, the watershed support area was delineated and a suite of landscape variables was calculated. There were 244 sites with its supported watershed. A total of 46 landscape metrics was generated per each watershed. Measured total phosphorous (TP), total ammonia (TAM), and *E. coli* only existed in 18, 6, and 15 sites, respectively. Landscape metrics were for year 2000 and surface water constituents were averaged over a period of 1997-2002. Each of the surface water constituents from the above sites was used in PLS to predict for the remaining from the 244 sites.

2.3. STATISTICAL METHODOLOGY

PLS is a multivariate analysis technique that permits analysis and prediction for data sets with missing values, with collinearity and with a relatively small number of observations (see references in Nash et al., 2005).

In the PLS analyses, both data sets (e.g., water and landscape variables) are first centered and scaled. A linear combination is composed of the independent variables ($T = L_o W$; T is the score and W is weight) forming a number of orthogonal latent variables [T] that are less in number (dimensions) than that of the original landscape variables. The linear combination in [T] is formed so that the covariance between [T] and the linear composition of the dependent variables are maximized ($T \& U$; $U = B_o V$; U is the score and V is weight). Prediction of both water and landscape data will be via regression on the common latent variables (T). Modeling and prediction in PLS, therefore, is not solely based on the conditional distribution of the predictors (water variables) in the presence of independent variables (landscape variables); instead it accounts for both landscape and water together through [T] (see references in Nash et al., 2005).

PLS produces n-1 factors, with each factor containing a pair of scores (T_i, U_i). Linear combinations on each data set are called factors. For example, if the number of sites (observations) is 89, then 88 factors will be produced. Not all of these factors are significant using the Cross Validation (CV) method; only the significant factors are used in the final model. The fitted models are tested using the test data sets and the predicted values are compared with that of observed using PRESS (Predictive Residual Sum of Square) to assess the predictive ability of the model. Root means PRESS and its significant level (the lower the value, the better the model is) will be used in the final model.

After defining the significant PLS factor, scores, weights, and VIP (Variable Influence on Projection) are used to examine the strength of the relationship, irregularities, and the contribution of the independent variable (landscape) in the model. It was indicated VIP values of less than 0.8 are considered to be small. The quality of the model was determined by examining the residuals for both the response and the landscape variables for any possible outliers. SAS was used for all statistical analyses.

3. Results

TP PLS model resulted in one significant factor explaining 59% of the variability in the TP (see table). The most significant contributors are the watershed percent barren and stream density. While the stream density relates inversely with TP, percent barren enhances TP in surface water. The forest-related variables contribute equally with a negative effect on the TP. Urban enhanced TP but mostly within the proximity (Rurb0) of the sampling site.

TAM PLS model resulted in one significant factor explaining 93% of the variability in the TAM (see table). Riparian and natural within all distances have negative effect on TAM, whereas urban has a positive effect.

E. coli PLS model resulted in two significant factors explaining 81% of the variability in the *E. coli* population (see table). Urban in riparian within area of 0m or more of the sampling site did enhance *E. coli*.

The prediction of the constituents in the 244 watersheds (from a small field-based data sample) was used to visualize the joint behavior of the predicted TP, TAM, and *E. coli* in surface water of the Upper White River (Figure 2).

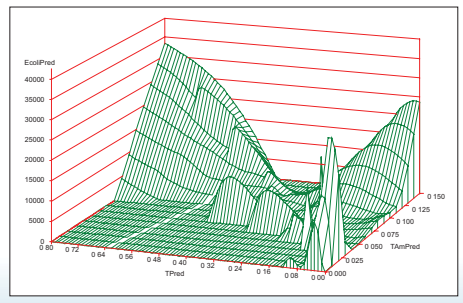

Figure 2. Three-dimensional plot of predicted TAM, TP, and *E. coli* cell counts among 244 subwatersheds in the Upper White River region of the Ozarks.

Landscape Metrics	TP Coefficient	TP VIP	TAM Coefficient	TAM VIP	*E. coli* Coefficient	*E. coli* VIP
Intercept	-2.10634		0.07867		0.92667	
F_c2a210	0.00368	0.87681				
F_mdcp	0.00249	0.92441				
F_plgp					-0.00052	0.82952
Pfor	-0.00319	1.04904			-0.00206	0.94050
Rfor0	-0.00334	0.95938			-0.00002	0.94904
Rfor30	-0.00327	0.98311			-0.00078	0.95368
Rfor120	-0.00326	1.02739			-0.00139	0.95435
Rnat0			-0.00027	0.91965	-0.00002	0.94904
Rnat30			-0.00025	0.92898	-0.00078	0.95368
Rnat120			-0.00023	0.94123	-0.00139	0.95435
Purb	**0.00270**	**1.03313**	**0.00031**	**1.06739**	**0.00559**	**1.10603**
Rurb0	0.00397	1.01606			0.00722	1.05721
Rurb30	**0.00378**	**1.02533**	**0.00042**	**1.07047**	**0.00676**	**1.05577**
Rurb120			0.00036	1.06935	0.00595	1.05542
Rhum0					0.00002	0.94904
Rhum30					0.00078	0.95368
Rhum120					0.00139	0.95435
Pctia_rd	0.00512	1.01764			0.01028	1.07970
Rddens	0.03096	1.05578			0.06105	1.10793
Pmbar	0.68643	0.87789			2.98807	1.26598
Rmbar0					0.62089	0.89036
Rmbar30					0.87270	0.88389
Rmbar120					2.58041	1.31004
Strmdens	-0.37189	1.16001			-0.94186	1.24181
Elevmean					0.00006	0.93605
Elevmin					0.00007	1.40485
Landarea(km^2)					-0.00048	0.96849
Number of Factors	1		1		2	
% Variation	59		93		81	

Coefficients of the non-centered value of landscape metrics to predict the ln(TP), TAM, and ln(*E. coli*). Number of significant PLS factors and percent variation explained by PLS for the responses are in the last two rows.

4. Discussion and Conclusion

The results indicate PLS may prove to be a valuable statistical analysis tool for ecological studies. The PLS methodology is less sensitive to the limitations than other statistical methods. The joint behavior of TP and TAM as related with *E. coli* (Figure 2) was not possible using the measurements from the study area sites (18, 6, and 15 sites for TP, TAM, and *E. coli*, respectively), but it was overcome by prediction from the PLS model for the 244 sites. Hence, further analyses and comparisons within and between the four groups (high TP-high *E. coli*, low TP-high *E. coli*, low TP-high *E. coli*, and moderate TP, TAM, and *E. coli*) may reveal the spatial characteristics setting for watersheds and their effect on surface water quality.

The model results may help landscape ecologists produce indicators of surface water condition, such that unique combinations of these indicators can be used to infer the potential cause(s) and origin(s) of nonpoint pollution, which may lead to eutrophication in aquatic ecosystems, the loss of aquatic ecosystem function, and the injury of humans that consume from (or recreate in) the aquatic resources of the Ozarks. The PLS results discussed in this poster are actively being used to prioritize subwatersheds in the Ozarks for watershed management activities.

Reference

Nash, M.S., D.J. Chaloud and R.D. Lopez. 2005. Application of Canonical Correlation Analysis and Partial Least Square Analyses in Landscape Ecology. EPA/600/X-05/004.