Application of Partial Least Square (PLS) Regression to Determine Landscape-Scale Aquatic Resources Vulnerability in the Ozark Mountains

Maliha S. Nash and Ricardo D. Lopez U.S. EPA, Office of Research and Development, National Exposure Research Laboratory, Environmental Sciences Division, P.O. Box 93478, Las Vegas, Nevada 89193-3478

Introduction

Partial least squares (PLS) analysis offers a number of advantages over the more traditionally used regression analyses applied in landscape ecology, particularly for determining the associations among multiple constituents of surface water and landscape configuration. Common data problems encountered during landscape-ecological analyses may include small sample sizes, missing data values among sampled areas, a large number of predictor variables, correlated variables, and high noise-to-signal relationships. PLS attempts to account for the above data problems, by building a robust association model. We utilized PLS to predict *in situ* surface water *Escherichia coli* (*E. coli*) bacterial counts in the Upper White River from the associated landscape-ecological metrics in the Ozark Mountains (southwestern Missouri and northwestern Arkansas, USA). The amount of variability in *E. coli* counts was explained by each PLS model and reflects the composition of the contributing landscape among the watersheds analyzed. The predicted values and their confidence intervals explain how land cover type and configuration, and land use may affect the abundance of *E. coli* in surface waters of the Upper White River region of the Ozark Mountains.

1 - Site Description

The study area is a 21,848 square kilometer area of land that encompasses the headwaters of the White River, and generally the Ozark Mountains (Figure 1). The study area contains a mix of pasture and other agriculture (e.g., poultry production facilities, cattle operations, and hay operations), forest, and urban land cover, as well as several large reservoirs (Figure 2). The White River originates in northwestern Arkansas and flows through southwestern Missouri and north-central Arkansas. The White River descends from the Ozark Mountains into Arkansas' agricultural plain where it meanders to its confluence with the Mississippi River (not shown in Figure 1).

Figure 1. The study area is in the Upper White River study area (21,848 km²) in the Ozarks of Missouri and Arkansas, shown as four separate 8-digit U.S. Geological Survey hydrologic unit codes (HUCs).

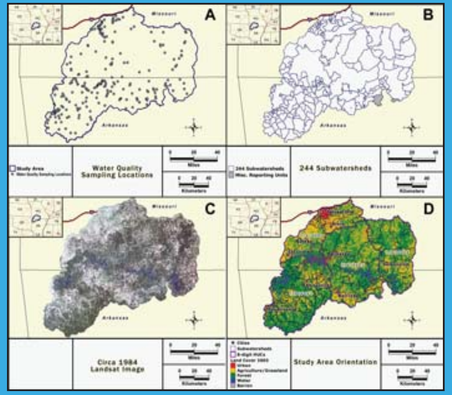


Figure 2. The Upper White River study area is in the Ozarks of Missouri and Arkansas, where 244 water quality sampling locations were sampled (A) and used as "pour points," from which 244 contributing sub-watersheds were delineated (B). A combination of multiple Landsat Thematic Mapper imagery (C) and 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 - Data Description

Water Bacterial Variable

Escherichia coli in surface water measurements from 1997 to 2002 were compiled from U.S. Geological Survey and state agency data sets from 244 possible stream sample locations. *E. coli* is a species of fecal coliform bacteria that is specific to fecal material from humans and other mammals and birds. We selected *E. coli* as a surface water response parameter because EPA recommends it as one of the important indicators of health risk from water contact in recreational waters (U.S. EPA, 1997). Sources of *E. coli* contamination in surface water include municipal wastewater treatment plants, ineffective septic systems, domestic animal manure, wild animal feces, and storm water runoff (Lory, 1999). The water bacterial data (Y) used in this analysis were the abundance of *E. coli* counts in surface water. The data were used as response variables in the PLS statistical model development.

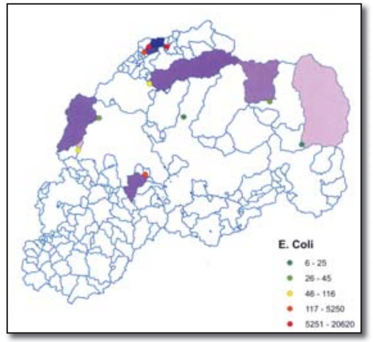


Figure 3. The non-nested watersheds with *E. coli* sampling points (n=10) that were used in the Partial Least Squares analyses. Elevated *E. coli* bacterial counts were positively correlated with landscape metrics that are indicators of human activities.

Landscape Variables

A total of 30 landscape variables (see table for variable description) used in this analysis are derived from available digital data sets in a geographic information system (GIS). Most of the landscape variables were calculated using the delineated drainage area (watershed) above the field sampling point as the base unit. These variables represent the percent forest, urban, human, agriculture and barren areas within sub-watersheds and within different proximities to streams (i.e., within riparian zones). Other variables such as elevation, stream density, road density and impervious layer were also included in the model.

Abbreviation	Landscape Metric Description
$Agtsl_3$	Percent agriculture (including all cropland and pastureland/grassland) Land-cover on slopes greater than 3%
Agt_0	Percent agriculture (including all cropland and pastureland/grassland) Land-cover adjacent to streams and rivers
Agt_{30}	Percent agriculture (including all cropland and pastureland/grassland) Land-cover within 30 meters of streams and rivers
Agt_{120}	Percent agriculture (including all cropland and pastureland/grassland) Land-cover within 120 meters of streams and rivers
Agt	Percent agriculture (including all cropland and pastureland/grassland) Land-cover within sub-watershed
For_0	Percent riparian forest land-cover adjacent to streams and rivers
For_{30}	Percent riparian forest land-cover within 30 meters of streams and rivers
For_{120}	Percent riparian forest land-cover within 120 meters of streams and rivers
For	Percent forest land-cover within the sub-watershed
Nat_0	Percent natural land-cover adjacent to streams and rivers
Nat_{30}	Percent natural land-cover within 30 meters of streams and rivers
Nat_{120}	Percent natural land-cover within 120 meters of streams and rivers
Hum_0	Agt_0+Urb_0
Hum_{30}	$Agt_{30}+Urb_{30}$
Hum_{120}	$Agt_{120}+Urb_{120}$
$Mbar_0$	Percent barren land-cover adjacent to streams and rivers
$Mbar_{30}$	Percent barren land-cover within 30 meters of streams and rivers
$Mbar_{120}$	Percent barren land-cover within 120 meters of streams and rivers
Mbar	Percent barren land-cover within sub-watershed
Urb_0	Percent urban land-cover adjacent to streams and rivers
Urb_{30}	Percent urban land-cover within 30 meters of streams and rivers
Urb_{120}	Percent urban land-cover within 120 meters of streams and rivers
Urb	Percent urban land-cover within sub-watershed
Elevmin	Minimum topographic elevation
Pctia_rd	Percent impervious surfaces based upon roads
Rddens	Total road density
F_plgp	Percent of entire sub-watershed comprised by largest patch of forest
Strmlen	Total stream length
Strmdens	Total stream density
Flargest	Area of largest forest patch

3 - Statistical Methods Overview

The 30 landscape variables were related to the *E. coli* count in the PLS modeling procedures. The PLS model was built using a 'non-nested watershed' approach (n=10; Figure 3) and the *E. coli* values were log-transformed. Cross validation (i.e., holding one value out) was used to finalize the model, which retained the significant model factors (P > 0.05, van der Voet, 1994). The relative importance and coefficient values of each of the 30 predictor (X) variables were analyzed for their relationships with, and prediction of, surface water *E. coli* counts (Figure 4). Based on Figure 4, predictors with small coefficient and VIP < 0.8 can be removed and a new model can be built. The reduced models (23 landscape variables) still have one significant factor with a minimum root mean PRESS = 0.4969, but have a lower percent variation accounted for by PLS (85.2%). The VIP values for all 23 variables > 0.8.

4 - PLS Step-by-Step

1. Center and scale each of the response (Y) and predictor (X) variables, Y^o and X^o, respectively.
2. Construct linear combinations of the predictors as:

   $$\delta(score) = X^o \omega(weight)$$ Scores are orthogonal

3. Construct linear combinations of the response as: $\mu = Y^o \nu$
4. Verify the linear combination in (2) has maximum covariance with the response linear combination in (3); in addition constraints $\omega^T \omega = 1$ and $\delta^T \delta = 1$ should be met. $(\delta' \mu)$
5. Predict for both Y^o and X^o by regression on (δscores):

   $$\hat{X}^o = \delta L'_x$$
   $$\hat{Y}^o = \delta L'_y$$

   where $L'_x\ (= (\delta'\delta)^{-1}\delta' X^o)$ and $L'_x (= (\delta'\delta)^{-1}\delta' Y^o)$ are the X- and Y- loadings

6. The above steps are for constructing the first PLS factor
7. Residuals for each X and Y are produced as:

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

8. The second factor is constructed by applying steps 1 through 5 to the residual (7); additional factors are constructed by repeating this process for each residual until the X matrix becomes null. Weights are the contribution of each the predictors in X to the PLS factor. The scores are the regression coefficients of the variables in X and Y regressed upon the various variables in δ and represent how the different manifest variables are related to the scores δ. The scores are sometimes thought of as latent unobservable variables
9. We also endeavored to find the statistically significance of predictors and the reliability of predicted response values.
   – To identify the role of predictors in explanatory power on the response variable, statistical significance of predictor coefficients was assessed using the 95% confidence interval (CI) using bootstrapping. If the confidence interval of a coefficient crosses the zero value, it implies the nonsignificant contribution of that predictor.
   – The reliability of the predicted values can also be assessed. We used the 5th and 95th percentile for the predicted response (*E. coli*).

5 - Statistical Output

Since the reduced model did not improve in the percent variability explained, we used the initial model to predict the *E. coli* in other watersheds.

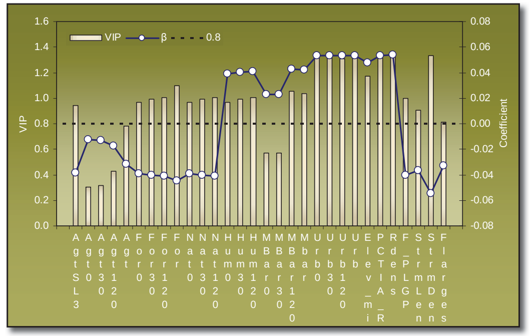


Figure 4. Demonstration of the coefficient (β) and relative importance [variable influence of projection (VIP)] values for the 30 landscape variables in the PLS model. Number of significant factors = 1; Minimum root mean PRESS = 0.4123; Percent variation accounted for by Partial Least Square factors for the dependent variable = 89.1%. Coefficient estimated for centered and scaled data.

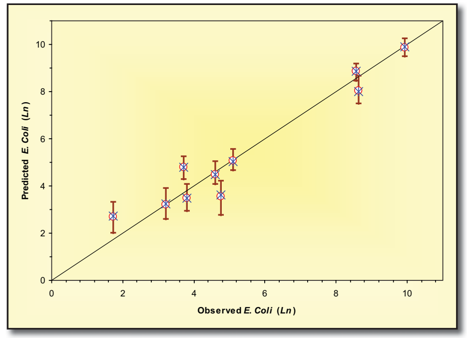
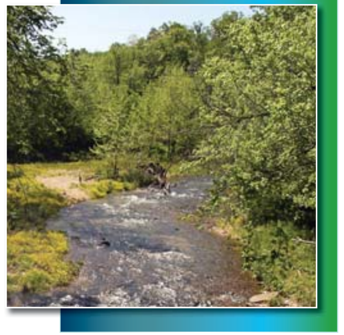


Figure 5. The observed *E. coli vs.* the predicted boot-strapped mean (red open circle), median (star), and the 5th and 95th percentiles in respect with the 1:1 relationships.



Results and Discussion

Despite a relatively small sample size (n=10), PLS permitted valid analyses of the Ozarks data, where other multivariate analyses provide fewer options. The analyses revealed that different landscape variables likely affect surface water (bacteriological) biota, based upon spatially explicit parameters. The role of urban and human activities enhanced the level of *E. coli* counts but more so within proximity of the stream (β in Figure 4: Urb_0, Urb_{30} > Urb_{120} > Urb_ws), than with the sub-watershed as a whole. While a decrease in slope within the sub-watershed enhanced the *E. coli* count, stream density and stream length resulted in a decrease in *E. coli* counts, perhaps as a result of a dilution effect. Overall, an increase in the amount of forest, whether by percentage or by forest patch size within a sub-watershed, decreased *E. coli* counts, likely as a result of either the physical impediment to surface flow of bacteriological contaminants, by forest vegetation, or biological interactions within those forested areas, or by lack of inputs. Further investigation of the effects of riparian vegetation on the amelioration of bacteriological contaminant in rivers and streams of the Ozarks is needed to verify these models.

The significant role of the landscape variables into prediction of *E. coli* can be assessed by their confidence intervals (Figure 6). While natural, forest, stream length and $AgSl_3$ (agriculture on slopes greater than 3 percent) are (confidence interval does not cross zero) negatively associated with the level of surface water *E. coli*, the presence of humans in the landscape is a likely contributor to an increase in *E. coli* counts. Urban and agriculture metrics are crossing the zero value, denoting their non-significance. The effect of agriculture on *E. coli* counts is higher within closer proximities to surface water, i.e., decreases with greater distances from agriculture.

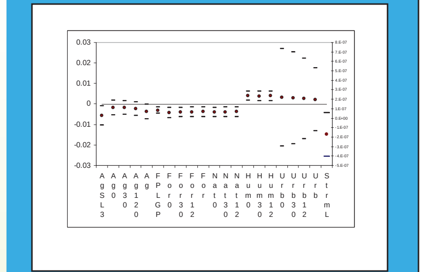


Figure 6. The estimated value for the coefficient of each predictor and its 95% confidence interval from bootstrap method. Right y-axis is for Strm Length.

Figure 7. The level of agreement between the predicted (polygon) and observed (closed circle) *E. coli*. The synchronization of the colors of the polygon with that of the closed circle denote the level of agreement.

However, urban has an enhancing role in *E. coli* with an uncertainty range that is wider and overlapping all of the remaining metrics. Although the urban confidence intervals are crossing the zero value, they are overlapping other landscape metric's confidence intervals, indicating that there are confounding (correlation) relationships between them. The prediction of surface water *E. coli* counts from PLS and proximity to observed values are presented as a map (Figure 7), showing the agreement between the predicted value (color of the polygon) with that of the pour point.

PLS analyses offer a number of advantages over the more traditionally used regression analyses. PLS offers a valid statistical model when the number of samples is small, compared to the number of variables, and when there is a high degree of collinearity between predictors as well as responses. Additionally, the prediction error in PLS is smaller than in other multivariate methods. The advantages of PLS makes it an attractive statistical tool for development of landscape ecology models. Available real-world data sets for the Ozarks provided a realistic ecological data set to initially develop this tool for such studies. These data sets contain all of the limitations that hinder use of other multivariate statistics, i.e., small number of sampling sites, large number of variables, several different types of field-collected surface water data and remote sensing-derived landscape characteristics data. Currently, we are studying other approaches (e.g., Morris, 2009) in determining the confidence intervals for the predicted response variable.

References

Nash, MS, and Chaloud, D. 2002. Multivariate Analyses (Canonical Correlation Analysis and Partial Least Square, PLS) to Model and Assess the Association of Landscape Metrics to Surface Water Chemical and Biological Properties using Savannah River Basin Data. EPA/600/R-02/091.

Lopez, RD, Nash MS, Heggem DT, and Ebert DW. 2008. Watershed Vulnerability Predictions for the Ozarks using Landscape Metrics. Journal of Environmental Quality. 37(5): 1769-1780.

van der Voet, H. Comparing the Predictive Accuracy of Models using a Simple Randomization Test, Chemometric and Intelligent Laboratory Systems, 25, 313 (1994).

Morris, RE, Hammond, MH, Cramer, JA, Johnson, KJ, Giordano, BC,. Kramer, KE, and Rose-Pehrsson, SL. 2009. Rapid Fuel Quality Surveillance through Chemometric Modeling of Near-infrared Spectra. Energy and Fuel 23: 1610-1618.

24872leb10.Nash-VulnerabilityOzarkMtns