Journal of Mathematics and Statistics 1 (1): 29-34, 2005
ISSN 1549-3644

Association of Landscape Metrics to Surface Water Biology in the Savannah River Basin

[1]Maliha S. Nash, [1]Deborah J. Chaloud and [2]Susan E. Franson
[1]US Environment Protection Agency, P.O. Box 93478, Las Vegas NV 89193-3478,USA
[2]US Environment Protection Agency, 26 W. Martin Luther King, Cincinnati OH 45268,USA

Abstract: Surface water quality for the Savannah River basin was assessed using water biology and landscape metrics. Two multivariate analyses, partial least square and canonical correlation, were used to describe how the structural variation in landscape metrics may affect surface water biology and to define the key landscape variable(s) that contribute the most to variation in surface water quality. The results showed that the key landscape metrics in this study area were: percent forest, percent of total area in agriculture (row crops + pasture) on moderately erodible soils, percent of total area with slopes greater than 3% and stream density. The first two canonical variates describe the linear combinations of the two data sets (r = 0.74 and 0.63), weighted mainly by percent of total area with slopes greater than 3%, taxa richness of sensitive insects to pollution (EPT; Ephemeroptera-Plecoptera-Trichoptera Index), algal growth potential and percent of total area in agriculture on moderately erodible soils.

Key words: Savannah River Basin, Surface Water Biology, Landscape Metrics, Canonical Correlation

INTRODUCTION

A current interest in landscape ecology is in determining the relationships between landscape metrics (forest percent, road density) and properties of surface water quality. Surface water quality is related to conditions in the surrounding environment. Clearing vegetation, for example, exposes soils to water and wind erosion that enhances sediment transport to surface waters. Nutrients from agricultural sources eventually drain into surface water causing elevated levels of nutrients which can harm aquatic biota. Measurements of surface water and landscape metrics from many sites, therefore, are collected to analyze and explore relationships in an area using a number of statistical methods. Univariate- and multiple-regression analyses have frequently been used to relate water nutrient concentrations to selected landscape metrics [1-3]. These studies 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 simultaneously (landscape metrics and surface water biology) to describe their association and to explain their connection to their physical environment, multivariate analyses, such as canonical correlation [4, 5] should be used.

Canonical correlation analyses are used to describe association, including the association between vegetation species and environmental conditions [6, 7]. Results are used to describe the physical processes that lead to vegetation variation as a response to environmental conditions. Canonical correlation requires a relatively large number of observations (sampling sites) compared to the number of variables. In ecology, sample size is often small and the number of variables is frequently large. Therefore, we propose using another multivariate method, partial least square, PLS [5], to reduce the number of independent variables and retain those with the highest correlation to the response variables. The objectives of this study are to: (1) develop valid models for the relationship between landscape variables and surface water biological properties, (2) obtain information about which landscape variables that are important and how much of the biological properties that can be predicted, (3) interpret these models in order to obtain a better understanding and (4) to model for prediction for different ecoregion.

MATERIALS AND METHODS

The surface water data used in these analyses were provided by U.S. Environmental Protection Agency (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 [8]. Two data sets, surface water biology and landscape metrics were used in the analyses (Table 1). Total number of sites used for this study were 85.

Table 1: Water Biology and Landscape Metrics Used in the Canonical Correlation Analyses

Variable	Description
Water Biology	
Hab	Macroinvertebrate habitat.
EPT	Taxa richness of sensitive insects to pollution; EPT stands for Ephemeroptera-Plecoptera-Trichoptera Index. These insects are correlated with good water quality [19] based on 100-organism subsample, non-impacted (>10), slightly impacted (6-10), moderately impacted (2-5) and severely impacted (0-1).
Rich	Macroinvertebrate richness, species richness (SPP). Total number of species in a sample. Conditions of areas are classified as: non impacted (>26), slightly impacted (19-26), moderately impacted (11-18), severely impacted (<11).
AGPT	Algal Growth Potential Test
IBI	Fish Index of Biotic Integrity [20]
Landscape Metrics	Percent area in a HUC
Pct_for	Percentage of total MRLC landcover in forest types
Ag_mod	Percent of total area in agriculture (row crops + pasture) on moderately erodible soils (STATSGO K-factor 0.2 and<0.4)
Slope3	Percent of total area with slopes greater than 3%
Strmden	Stream density as total length of streams from USGS TIGER data

Study Site: The Multi Resolution Land Characteristics Consortium (MRLC) land cover/land use data for Savannah River Basin (http://www.epa.gov/nerlesd1/land-sci/savannah.htm) 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 wetland areas along the river. The city of Savannah can be seen near the outlet of the river to the Atlantic Ocean. The spatial patterns seen in the land cover correspond closely to the four ecoregions: Blue Ridge, Piedmont, Coastal Plain and Atlantic Coastal Plain. Only the first three ecoregions were used in these analyses due to the paucity of water samples for the Atlantic Coastal Plain.

Statistical Analysis: To assess the relationships between the two data sets of water biology and landscape metrics, we used partial least square (PLS) and canonical correlation analyses. Detailed descriptions of these two methods are given in [5]. A brief description of PLS is summarized below for the reader. Partial least square projection of latent structures is a multivariate method and it is specifically close to canonical correlation. Partial least square is widely used in chemometrics for quantitative structure property relationships research to describe how structural variations in chemical compounds affect biological activity. Also, PLS predicts the chemical form from spectroscopy readings, where several hundred wavelengths and a smaller number of chemical samples [9] is the norm in chemometric analyses. For example, in quantification of molecular modeling, a large number of independent variables (>1000) are normally obtained with respect to the number of samples (10 to 100).

We used both data sets (5 biological variables and 26 landscape metrics) from 85 sampling sites simultaneously in PLS. One output from PLS is variable influence on projection (VIP) that can be used to select the most important landscape variables as related to the surface water biology for further analysis (Table 2). A VIP value of less than 0.8 is considered to be small contributor [10]. Landscape metrics with VIP >1 were selected. When many metrics (VIP >1) describe a similar attribute (Ag_mod, Ag_slp, Ag_slp_mod), only one of these metrics was chosen (Ag_mod). Based on the relative importance from PLS, collinearity and multinormality [5], we selected four landscape metrics; Ag_mod, Pct_for, Strmden and slope3. Five water biology variables; algal growth potential test (AGPT), macroinvertebrate habitat (hab), taxa richness of sensitive insects to pollution (EPT), macroinvertebrate richness (Rich) and fish Index of Biotic Integrity (IBI) were related to the four landscape metrics using canonical correlation analyses to describe association and variability structure between the two data sets. In addition, each water biology variable was modeled using the landscape metrics as the independent variables in step-wise multiple regression. Additional landscape metrics; total length of roads in watershed (TotRoadWS) and percent of total area with highly erodible soils (soil_er); were added to the regression model knowing that these variables have a different effect on surface water based on the geographical locations (i.e. ecoregion). Hence the strength of the model may be improved.

There were a number of missing values in the water biology data set. The procedure we used [11], deletes

Table 2: PLS Regression Coefficient and Variable Influence on Projection (VIP) Values (5 Biota, 26 Landscape Metrics and 85 Sampling Sites). Asterisk Denotes the Landscape Metrics Used in the Canonical Correlation Analysis

Predictor	AGPT	EPT	IBI	Hab	Rich	VIP
Ag_hi	-0.004	0.005	-0.001	0.003	0.003	0.11
Ag_slp_hi	-0.005	0.006	-0.001	0.003	0.004	0.14
Ag_mod*	0.054	-0.065	0.012	-0.039	-0.047	1.50
Ag_slp	0.044	-0.053	0.010	-0.032	-0.038	1.23
Ag_slp_mod	0.045	-0.054	0.001	-0.032	-0.039	1.26
Bar_slp_hi	-0.006	0.007	-0.002	0.004	0.005	0.17
Bar_slp_mod	-0.009	0.010	0.011	0.006	0.008	0.24
Crop_slp	0.047	-0.057	0.011	-0.034	-0.041	1.33
Crop_slp_mod	0.050	-0.061	0.005	-0.036	-0.044	1.42
Past_slp	0.038	-0.046	0.009	-0.027	-0.033	1.06
Pct_bar	-0.010	0.012	-0.002	0.007	-0.008	0.27
Pct_crop	0.032	-0.039	0.007	-0.023	-0.028	0.90
Pct_for*	-0.054	0.065	-0.012	0.039	0.047	1.51
Pct_past	0.046	-0.056	0.011	-0.033	-0.040	1.30
Pct_urb	0.022	-0.026	-0.005	-0.016	-0.019	0.61
Pct_wet	-0.008	0.009	0.002	0.005	0.007	0.21
Pct_wtr	0.025	-0.031	0.006	-0.018	-0.022	0.71
PwrPipTL	-0.005	-0.007	0.001	-0.004	-0.005	0.15
Slope3*	-0.041	0.049	-0.009	0.029	0.035	1.14
Slp_mod	-0.005	0.006	-0.001	0.003	0.004	0.13
Sd_slp	-0.062	0.075	-0.014	0.045	0.054	1.75
Mean_slp	-0.062	0.074	-0.014	0.044	0.054	1.73
Soil_er	0.034	-0.041	0.008	-0.025	-0.030	0.96
Strmden*	-0.036	0.044	-0.008	0.026	0.032	1.02
TotRoad30	0.008	-0.010	0.002	-0.006	-0.007	0.23
TotRoadWS	0.023	-0.028	0.005	-0.016	-0.020	0.64

all records containing a missing value, resulting in a low number of observations. We used PLS to estimate for those missing water biology values from landscape variables that were not to be used in the canonical correlation analyses. Elevation, mean slope and standard deviation in slopes were used for estimation of the missing values. Canonical correlation analysis produces a number of canonical variables, each consisting of pairs of linear combinations (canonical variate) of landscape and biology. Values of the coefficients of each of the original variables in its canonical variate are determined so that the correlation between pairs of the first canonical variate (Bio1 and LS1) is maximized (hereafter, we will use Bio1-LS1 and Bio2-LS2 to denote the water biology and landscape pairs in the first and second canonical variates, respectively). The first pair of canonical variates always has the highest canonical correlation (r_k). The number of pairs (k) produced is equal to the number of variables in the smaller data set. In this study we have a total of four canonical pairs. We used canonical correlation, square canonical correlation and canonical coefficient for interpreting and assessing the variance shared between the linear combination of both data sets. A coefficient value of each of the original variables in the canonical variate reflects the contribution of each in presence of others (Fig. 1). The linear compositions of the two data sets (original variables) were built to give the maximum correlation between the two canonical variates. Hence, multivariate context is more evident in coefficients than that in correlation of the original variables with its canonical variate. The number of significant canonical variates will constitute the final model for biota as related to surrounding landscape conditions. We used SAS for all the statistical analyses; proc PLS, proc cancor and proc reg [11]. Theory and development of canonical correlation are well stated in many references [12-17].

We used canonical correlation analyses to test the hypotheses of: (1) no linear relationships (Ho: Cov(yx) = 0) between the surface water biology and landscape metrics, (2) whether any observed linear relationships are significant and (3) define the rank of the model, i.e. the number of significant canonical variates. We used a significance level of 0.05 for all tests.

RESULTS AND DISCUSSION

Surface water biology indicators and landscape metrics were dependent and exhibited significant relationships

Table 3: Canonical Correlation (R_k), Canonical R-Square (R^2_k), Percentage Trace (% Trace = R^2_k *100 / R^2_k) and Cumulative Trace (%) of Total (Standardized) Sample Variance Explained by Each Canonical Variate and the Probability of Exceeding the Critical Value (F)

k	r_k	r^2_k	Trace %	Trace Cum. (%)	P>F
1	0.742	0.551	74.45	74.45	<0.0001
2	0.625	0.391	18.79	93.24	<0.0001
3	0.220	0.048	6.48	99.73	0.6642
4	0.047	0.002	0.27	100.00	0.9167

Table 4: Coefficients and R^2 for Significant (P<0.05) Relationships Between Surface Water Biology (Dependent Variables) and Landscape Metrics (Independent Variables) in Step-Wise Multiple Regression Analyses

		Landscape Metrics						
	Biology	Slope 3	Ag_mode	Strmden	Forest	Soil_er	TotRoasWS	R^2
All	AGPT	+		-				0.28
	EPT	+	-					0.40
	Rich					+		0.10
	Hab					-	-	0.28
Ecoregion								
Blue-	AGPT	-						0.3
Ridge	EPT				+			0.43
	Rich				+			0.24
Piedmont	AGPT				-			0.27
	IBI	-						0.09
	EPT	+			+	-		0.43
	Hab					-	-	0.28
Coastal	IBI						-	0.28
Plain	EPT						+	0.77

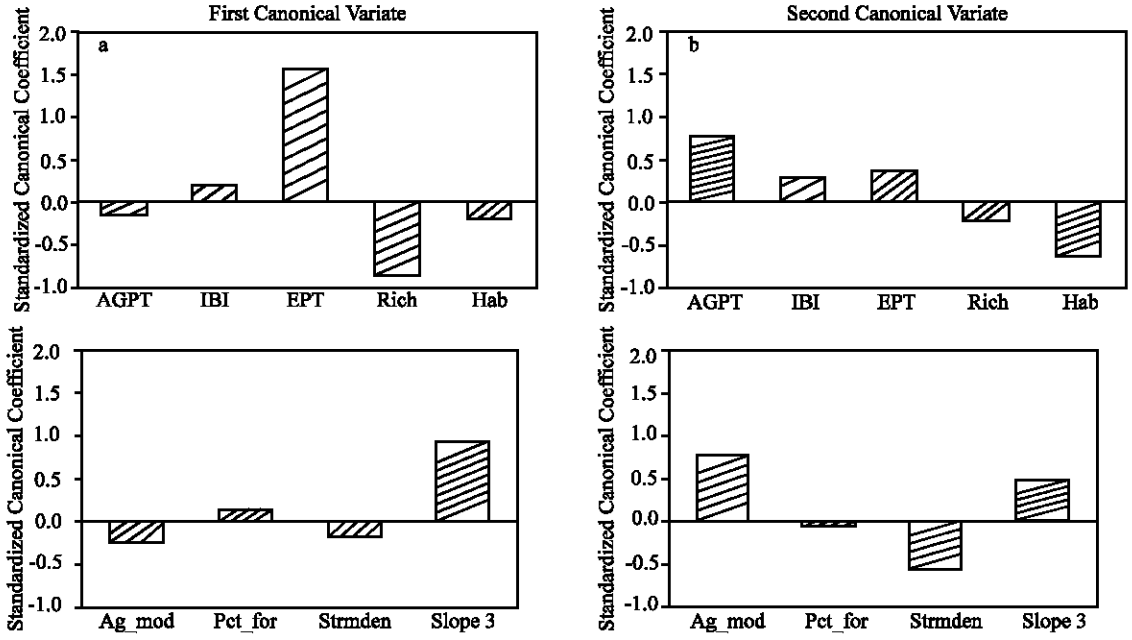

Fig. 1: Coefficient Values for the Two Significant Canonical Variates in the Model. Contributed Original Variables for the First Canonical Variate, (A) Slope 3 (Percent of Total Area with Slope >3%) and EPT (Taxa Richness of Sensitive Insects to Pollution) and Rich (Macroinvertebrate Species Richness), for the Second Canonical Variate (B) Ag_Mod (Percent of Total Area in Agriculture, Row Crops + Pature) on Moderately Erodible Soils, STATSGO (0.2 K-Factor<0.4) and AGPT (Algal Growth Potential Test)

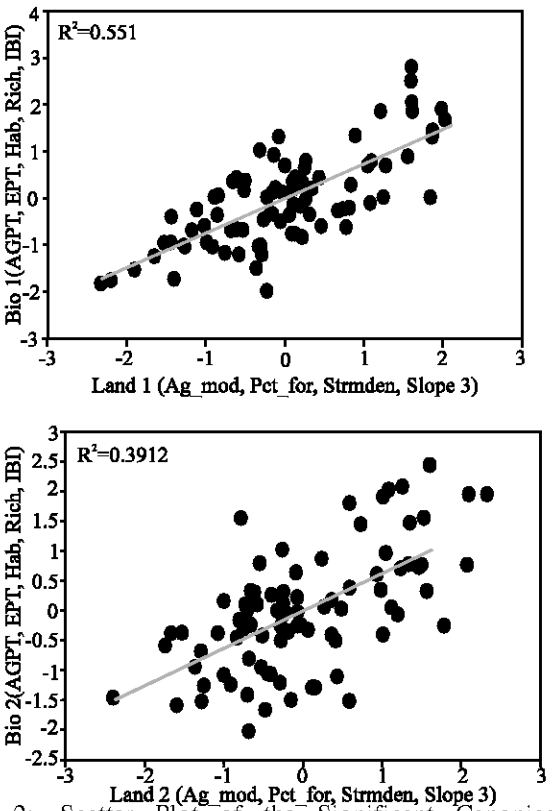

Fig. 2: Scatter Plot of the Significant Canonical Variates for the Landscape Metrics (Ag_mod, Pct_for, Strmden, Slope 3) and Water Biology (AGPT, EPT, Hab, IBI, Rich)

(Roy's test with $p<0.0001$). The canonical correlation values (r_k) for the first two canonical variates were significant ($p<0.05$; Table 3). The first canonical correlation ($r_k = 0.74$, Fig. 2) indicates a strong correlation between the linear composition of the water biology and landscape metrics in the first canonical variate, whereas, the second ($r_k=0.63$, Fig. 2) is on the high side of moderate correlation [18]. Squared canonical correlations (r^2_k; Table 3) express the amount of variability that is shared and explained by the two linear compositions of the water biology and landscape metrics. The first two canonical variates explained 0.55 and 0.39, respectively, of the variation as compared to 0.05 for the two remaining canonical variates. The sum of r^2_k (=0.992; Table 3) is a measure of the variance shared by all four pairs of canonical variates. The fitted model, which represents only the significant pairs, accounted for 93% (%Trace = (0.55+0.39) * 100 / 0.992; Table 3) of that overall shared variance. Percent trace can be used as an index of the fitted model quality that measures prediction adequacy.

While the first canonical variate was weighted mainly by slope3 and EPT, the second canonical variate was weighted primarily by Ag_mod and AGPT (Fig. 1). In the Savannah River basin, areas on higher slope provide more suitable environment for the sensitive macroinvertebrates. These areas of higher slope also have higher percent forest, additionally providing better environment for these insects (in a separate analysis, when slope3 was excluded, percent forest and EPT were weighted heavily on the second canonical variate ($p<0.01$)). As the stream density increased, richness increased and stream density had an effect in the same direction on both canonical variates.

Multiple Regressions per Ecoregion: To leverage a univariate relation of each of the water biology variables with many landscape metrics, multiple regression analyses with stepwise selection were applied. In addition to the landscape metrics that were used in canonical correlations, total length of roads in watersheds (TotRoadWs) and percent of total area with highly erodible soils (soil_er) were used in the multiple regressions. Regression analyses were done for the whole area and by ecoregion. R^2 values ranged from 0.10 to 0.43 (except for Coastal Plain; Table 4) indicating a range of weak to moderate relationships when an individual biota indicator was regressed with the landscape metrics. Not all water biology variables were significantly related to the landscape metrics. Relationships were variable between the whole study area and among the ecoregions. In the Blue Ridge ecoregion, percent forest and slope were the only landscape metrics related to EPT, Rich and AGPT. For the Coastal Plain, total road per watershed was the single landscape metric that related strongly with the IBI and EPT. Whereas, for the Piedmont ecoregion, slope, forest, percent of erodible soil and total road per watershed were related differently to four out of the five water biology variables. Relationships, therefore, were not similar across the ecoregions. The predictability of each of the biota from the landscape metrics, however, is not as strong when the water biology and landscape metrics were related simultaneously as in canonical correlation.

CONCLUSION

PLS was used mainly to reduce the number of landscape metrics (26) to the key landscape variables in this study area. These metrics were the percent of area with slope greater than 3%, percent forest, percent of area in agriculture (row crop + pasture) on moderately erodible soils and stream density. The utility of PLS here is in a multivariate context with a purpose similar to that of the stepwise selection in multiple regressions. Canonical correlation analyses resulted in a measure of the strength of the relationship expressed by the canonical correlation (r_k) between two sets of multiple variables, biological and landscape variables. The first two canonical correlation were significant indicating the existence of moderate to strong relationships between the two data sets and indicating they are not independent. The canonical model indicated three major contributing variables: the landscape metric, slope greater than 3% (Slope3), the water biology

variable EPT (an indicator of three macroinvertebrate genera) and Rich (an index of macroinvertebrate species richness). Within this model, the landscape variable percent area in agriculture (row crop + pasture) on moderately erodible soils was the second highest landscape contributor, with a positive relationship to the AGPT and a negative relationship with the macroinvertebrate habitat. This suggests agriculture on moderately erodible soils is contributing nutrients to the surface waters in the basin.

In summary, the above results indicated increased slope (indicating complex topography, generally occurring in the mountainous areas of the Savannah River Basin) is associated with increased macroinvertebrate quality, while the percentage of landcover in Ag_mod is associated with increased AGPT and declines in aquatic biota quality.

Although the univariate-multiple regression of each of the biota with that of landscape metrics was low-moderate, it is evident that the importance of any landscape metric is a function of its spatial location in the study area. The importance of forest and area with slope greater than 3% are the highest in Blue Ridge and decreased consistently across the adjacent ecoregions of the study area. Areas with slope greater than 3% and soil erodibility are the most important landscape variables in the Piedmont. The relative importance of percent agriculture on moderately erodible soil increased in the Coastal Plain. Most of the variability in AGPT was explained by the percent of total area in agriculture (row crops + pasture) on moderately erodible soils ($R^2 = 0.52$) but the relationship was not significant ($p>0.05$).

ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable input of the reviewers that made this report more comprehensive and easy to follow by the readers. The surface water biological datasets were provided by the U.S. Environmental Protection Agency (EPA) Region 4, Science and Ecosystem Support Division. The EPA, through its Office of Research and Development, funded the statistical research described here. It has been subjected to the Agency's review and approved for publication.

REFERENCES

1. Noy-Meir, I., 1974. Multivariate analysis of the semiarid vegetation in southeastern Australia. II Vegetation Catenae and environmental gradients. Australian J. Bot., 22: 115-140.
2. Jones, B.K., A.C. Neale, M.S. Nash, R.D. Van Remortel, J.D. Wickham, K.H. Riitters and R.V. O'Neil, 2001. Predicting Nutrient and Sediment Loadings to Streams from Landscape Metrics: A Multiple Watersheds study from the United States Mid-Atlantic Region. Landscape Ecol., 16: 301-312.
3. Mehaffey, M.H. T.G.Wade, M.S. Nash and C.M. Edmonds, 2003. Analysis of Land Cover and Water Quality in the New York Catskill-Delaware Basins, pp: 1327-1339.
4. Cumming, S. and P. Vernier, 2002. Statistical Models of Landscape Pattern Metrics, with Application to Regional Dynamic Forest Simulation. Landscape Ecol., 17: 433-444.
5. Nash, M.S. and D.J. Chaloud, 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. Office of Research and Development, Washington, D.C. USA., pp: 75.
6. Ter Braak, C.J.F., 1987. The Analysis of Vegetation-Environment Relationships by Canonical Correspondence Analysis. Vegetation, 69: 69-77.
7. 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, USA.
8. Chaloud, D.J., C.M. Edmond and D.T. Heggem, 2001. Savannah River Basin Landscape Analysis. EPA/600/R-01/069. Office of Research and Development , Washington, D.C. USA., pp: 47.
9. Helland, I. S., 1988. On the Structure of Partial Least Square Regression. Commun. Statist. Simula., 17: 581-607.
10. 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.
11. SAS, 1998. Stat User's Guide. *SAS* Inst. Inc. Cary, NC USA.
12. Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, USA.
13. Rencher, A.C., 1998. Multivariate Statistical Inferences and Applications. John Wiley and Sons' INC. New York, USA.
14. Hair, J.F., R.E. Anderson and R.L. Tatham, 1987. Multivariate Data Analysis with Readings. Macmillan Pub. Co., New York.
15. Gittins, R., 1985. Canonical Analysis: A Review With Application in Ecology. Springer Verlag, NewYork.
16. Thorndike, R.M., 1978. Correlation Procedure for Research. Gardner Press, New York.
17. Clark, D., 1975. Understanding Canonical Correlation Analysis. Geo Abstract Ltd. University of East Anglia, Norwich, NR4 7TJ.
18. Sheskin, D.J., 2000. Handbook of Parametric and NonParametric Statistical Procedures. 2nd Edn. Chapman and Hall/CRD, New York.
19. Lenat, D.R., 1987. The Macroinvertebrate Fauna of the Little River, North Carolina: Taxa List and Seasonal Trends. Archiv für Hydrobiologie, 110: 19-43.
20. Karr, J.R., 1981. Assessment of Biotic Integrity Using Fish Communities. Fisheries, 6: 21-27.