COEUR D'ALENE LAKE TOTAL PHOSPHORUS NUTRIENT CALCULATIONS FOR LAKE TRIBUTARIES

July 2016 update

Vasile Alexandru Suchar, Department of Statistical Science, College of Science, University of Idaho, 875 Perimeter Drive MS1104, Moscow, ID 83844-1104, USA, vasiles@uidaho.edu

I. OBJECTIVE

Revise the multivariate regression analyses between total phosphorus (TP) loading and land use/land characteristic parameters for Coeur d'Alene (CDA) Lake tributaries that were performed in March 2016 with updated data.

II. WORK SCOPE

Test the prior variable selection with the revised stream data. Conduct additional regression analyses with the updated loading data. Use the NHD24k stream data. Also, include stream slope in the variable matrix. Determine a revised best fit.

III. STATISTICAL METHODS

1. Variable importance selection with Random Forest
2. Multivariate Regression Analysis
3. Model comparison

The analysis was conducted in R-language. For more details on the methodology and R language, consult the following references (Fox, 2008; Liaw and Wiener, 2002; R Core Team, 2015)

IV. VARIABLE IMPORTANCE SELECTION WITH RANDOM FOREST

Since we have only 12 observations, only a maximum of nine variable can be included in the Multiple Regression Analysis. To select these nine variables, I used Random Forest Variance Importance for 10,000 simulations. For the new dataset the predictor variables selected based on their importance are as follow (first 21 variables):

1. Stream_length24k
2. Road_length
3. Developed_open
4. Wetlands_herb
5. Wetlands_total
6. Total_true_forest
7. True_shrub
8. Developed_low
9. Evergreen_forest
10. Total_true_open
11. Stream_density24k
12. Mixed_forest
13. Hay_pasture
14. Total_crops
15. Stream_gradient
16. Crops
17. Road_density
18. Deciduous_forest
19. Harvest_ssh
20. Wetlands_woody
21. Developed_mid

Based on the objections regarding the use of catchment area in the prior analysis, I eliminated it from the dataset altogether in the current analysis. Since wetlands sampling of the sample dataset may not representative of the overall basin, analysis was also conducted on a dataset without wetland variables.

1. Stream_length24k
2. Road_length
3. Developed_open
4. Developed_low
5. True_shrub
6. Evergreen_forest
7. Total_true_forest
8. Total_true_open
9. Stream_density24k
10. Mixed_forest
11. Hay_pasture
12. Total_crops
13. Crops
14. Road_density
15. Deciduous_forest
16. Stream_gradient
17. Developed_mid
18. Harvest_ssh
19. True_herb
20. Total_harvest_forest
21. Harvest_evergreen

V. MULTIVARIATE REGRESSION ANALYSIS

Several OLS approaches were requested:

1. OLS model with wetland variables in the dataset
2. OLS model without wetland variables in the dataset
3. OLS model with wetland variables in the dataset & with stream gradient in the model
4. OLS model without wetland variables in the dataset & with stream gradient in the model
5. OLS model with wetland variables in the dataset & with % developed in the model
6. OLS model without wetland variables in the dataset & with % developed in the model

1. OLS model with wetland variables in the dataset

Model selection steps: following Random Forest variable importance step, a regression analysis for the maximum number of variables allowed was performed. Variance-inflation factors (VIFs) were calculated for the model. VIF indicates the impact of collinearity on the precision of the regression coefficients estimates. Variables that causes increases in VIF values were eliminated. More variables were added to the model. The previous steps are repeated until the model included nine uncorrelated variables (note: nine is the maximum number of explanatory variables that can be considered in the OLS regression due to the small sample size). Lastly, model selection based on significance of explanatory variables, adjusted R-squared, and diagnostic plots was conducted.

The final model (Results 1) has five explanatory variables, all significant at 0.10 level, and an adjusted R-squared of 0.9277. The AIC, BIC, and RMSE values for the OLS model 1 are also in Results 1.

Model 1:
```
Tp_coeff ~ Wetlands_total + Total_true_forest +
           Developed_low + Stream_density24k + Stream_gradient
```

Diagnostics plots show no major violations of the regression assumptions (Results 1). Observation 3 has moderate residuals and leverage; thus, moderate influence, not quite in the danger zone, but a little bit high. As before, observation 10 has high residuals but no leverage.

2. OLS model without wetland variables in the dataset

In this case we considered two approaches:

Approach 1: the same steps as before were followed. Often, the multicollinearity problem is caused by two highly correlated variables. Usually the elimination from the model of the variable with

the higher VIF value solves the problem. In this case, when this situation was encountered, alternative models using either variables were considered. This resulted in eight such models, further reduced to two following the modeling selection procedures. This modeling approach should allow the users the flexibility to choose the model that includes the variables considered more reliable.

The eight models created were along the combinations of three pairs of highly correlated variables: `Evergreen_forest` and `Total_true_forest`, `Stream_length24k` and `Road_length`, `Developed_open` and `Developed_low`. These creates an opportunity for more selective future data collection in the area. Further model selection procedures eliminated completely the `Evergreen_forest`, `Total_true_forest`, `Stream_length24k` and `Road_length` from the final models. The presented models (Results 2 and 3) contained either `Developed_open` or `Developed_low`.

Model 2: `Tp_coeff ~ Developed_low + True_shrub + Stream_density24k`

The final model (Results 2) has three explanatory variables, all significant at 0.05 level, and an adjusted R-squared of 0.7269. The AIC, BIC, and RMSE values for the OLS model 2 are also in Results 2. Diagnostics plots show no major violations of the regression assumptions. In this case observation 3 may extent significant influence on the regression coefficients.

Model 3:
```
Tp_coeff ~ Developed_open + True_shrub + Stream_density24k +
    Total_crops
```

The final model (Results 3) has four explanatory variables, all significant at 0.10 level, and an adjusted R-squared of 0.813. The AIC, BIC, and RMSE values for the OLS model 3 are also in Results 3. Diagnostics plots show no major violations of the regression assumptions. Observation 3 has moderate residuals and leverage; thus, moderate influence, not quite in the danger zone, but a little bit high.

Approach 2: In situations when *prediction*, instead of *understanding* which particular variables really affect the dependent variable *and to which extent* is important, multicollinearity might not be the primary concern in model selection. In this approach, explanatory variables were selected based on their significance. Multicollinearity was addressed at the end of the model selection. As a warning, if two or more explanatory variables are highly correlated, sometimes the estimates of their coefficients may not be reliable. The resulting model may have higher predictive power, at the cost of lower understanding of the relative contribution of the variables in explaining the variability in the data.

Model 4:
```
Tp_coeff^2 ~ Developed_low + True_shrub + Stream_density24k +
    Hay_pasture + Deciduous_forest + Developed_mid +
    Harvest_ssh + True_herb
```

The final model (Results 4) has eight explanatory variables, all significant at 0.05 level, and an adjusted R-squared of 0.9922. The AIC, BIC, and RMSE values for the OLS model 4 are also in Results 4. Diagnostics plots show no major violations of the regression assumptions. Observation 4 has high residuals, but no leverage. Thus, no significant influence on the regression coefficients.

`Developed_low` and `Developed_mid` are highly correlated, and but since in this approach prediction is favored instead of understanding, we will consider both in the model.

3. OLS model with wetland variables in the dataset & with stream gradient in the model

Model 1 satisfies these criteria.

4. OLS model without wetland variables in the dataset & with stream gradient in the model

For this model, ***Approach 1,*** as described in ***2. OLS model without wetland variables in the dataset*** was used.

Model 5: `Tp_coeff ~ Stream_gradient + Developed_low + True_shrub + Stream_density24k`

The final model (Results 5) has four explanatory variables, all significant at 0.05 level, and an adjusted R-squared of 0.6889. The AIC, BIC, and RMSE values for the OLS model 5 are also in Results 5. Diagnostics plots show no major violations of the regression assumptions. In this case observation 3 may extent significant influence on the regression coefficients.

5. OLS model with wetland variables in the dataset & with % developed in the model

Model 1 satisfies these criteria.

6. OLS model without wetland variables in the dataset & with % developed in the model

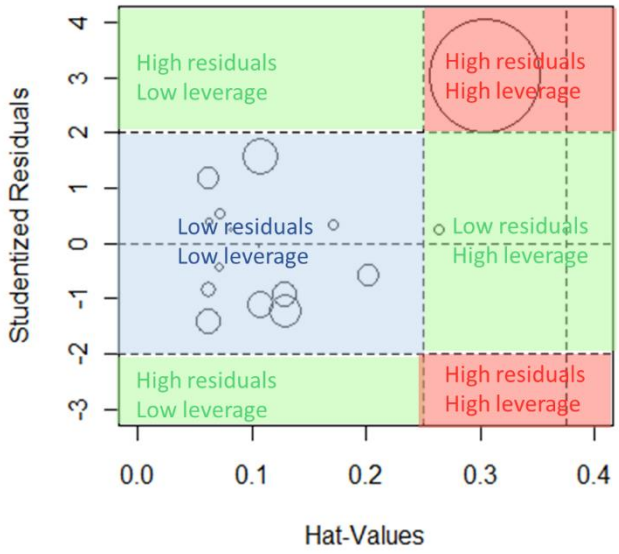
Models 2-4 satisfies these criteria.

VI. MODEL COMPARISON

The normality and homoscedasticity assumptions of the OLS regression analysis is satisfied by all models. Of course, the limited number of observations in the dataset, limits the effectivity of the diagnostics. Observation 3 may extent significant influence to regression coefficients in Models 2 and 5. The models should be used with caution.

Based on the model selection criteria, Model 4 followed by Model 1 seem to be the best supported by the data (Figure 1 and Table 1).

SUPPLEMENTARY INFORMATION



"BUBBLE PLOT" of studentized residuals, hat values, and Cook's distances (as areas of circles). The observations of concern are those with high residual and high leverage: they have potential to extent significant influence on the regression coefficients

MODEL SELECTION CRITERIA:

- Adjusted R-squared: takes in account the inflation of R-squared due to the number of regressors; the model with the highest adjusted R-squared value is the one most supported by the data.-
- The root mean square error (RMSE) is the square-root of the sum squared error for a model with p parameters including the intercept; it is often suggested to select the model with minimal RMSE.
- The Akaine Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are members of the general family of penalized model-fit statistics applicable to regression models fit by maximum likelihood. For both AIC and BIC, the model with the smallest value is the one best supported by the data.

REFERENCES

Fox, J., 2008. Applied regression analysis and generalized linear models, 2nd ed. SAGE Publications, Thousand Oaks, CA, USA 664 pp.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18-22.
R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.