Paper 140-2008

Data mining application of non-linear mixed modeling in water quality analysis

ANPALAKI J. RAGAVAN, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557

ABSTRACT

In regression analysis, non-linearity in fixed and random effects can adversely affect efficiency of regression parameter estimates. Successful non-linear time series modeling would improve regression parameter estimates and produce a richer notion of water quality than linear time series models allow. In addition multiple independent variables make each point in space a finite dimensional vector, non-linear in two dimensions jointly. The SAS/STAT procedure, NLMIXED fits non-linearity successfully in any time series using maximum likelihood-based methods.

In this data mining study spatial and temporal variations of total phosphorus concentration (TP) in Truckee River, Nevada, sampled monthly (from January 1997 to December 2004) over six sites were modeled as a function of soluble total phosphorus concentration (STP), stream flow (SF), seasonality (Summer), man-made intervention (X1), alkalinity, pH, temperature (Temp), dissolved organic carbon (DOC), and dissolved oxygen(DO) using the non-linear regression capabilities provided with the NLMIXED procedure in SAS® after successfully identifying non linearity in data. Likelihood ratio tests were conducted for model specification, and for tests of various hypotheses on individual cross sections. Results of parameter estimates, model diagnostics, and residual analyses were compared to that obtained from a linear mixed model fitted to the same data using PROC MIXED. Non linear model fitted data better. All independent variables influenced TP significantly ($p<0.0001$). Tests of cross sectional effects showed significant contributions of TP from all sites ($p<0.0001$) into Truckee River. Non-linearity in data can influence time series regression parameter estimates significantly.

KEY WORDS: PROC MIXED, PROC MI, PROC NLMIXED, Non-linearity, Maximum Likelihood

INTRODUCTION

Increasing urbanization, and population in Nevada in the past few years have created, increased concentrations of total phosphorus in the Truckee River. Water quality management of urban waters covers a broad spectrum of issues related to municipal, industrial and amenity irrigation practices. The natural cycle of phosphorus has been modified due to fertilizer use and sewage. Relationships of total phosphorus concentration to soil use and agricultural, domestic and industrial activities are expected to rise in the future. There is also tremendous uncertainty in pollution load estimation due to the associated spatial variability of pollutants in different catchments. Complexity in estimation of diffused pollutant concentration and load is a serious impediment to the design of efficient pollutant structures. Water management practices must be improved in Nevada, to guarantee improved quality of water of sustainable water bodies affected by development of urban and suburban areas. Determination of factors affecting or causing variation of phosphorus concentrations can provide a robust solution to quantify total phosphorus pollution in urban areas in Nevada.

Total phosphorus concentration (TP) is a random variable which must be handled probabilistically in space and time. The TP in a river is a function of several factors such as STP, SF, Summer, X1, DOC,DO, pH and water temperature. The degree of influence of the different factors on TP, need to be predicted to reflect different sources of the phosphorus loading into the river. The major focus of this research is predicting and modeling the non-linearity in the distribution of TP at the different sampling sites which enable designers to target TP concentration in the Truckee River accurately as close as possible to their source of origination. Non-linear time series modeling is an appropriate approach to analyze such data.

The classical Box and Jenkins models of time series analysis uses the Wold Representation (any covariance stationary time series can be expressed as moving average function of present and past innovations). This infinite moving average can nearly always be well approximated by low order autoregressive processes perhaps with some moving average components. Further, the dynamics of the time series can be directly 'read off' from the Wold Representation. Statistical models in which both fixed and random effects enter nonlinearly are becoming increasingly popular. Perhaps the greatest theoretical progress in time series analysis in the last ten years has been in the understanding of testing and modeling for nonlinearity. Nonlinear time series analysis raises the possibility of improving the power of parameter estimation and forecasting techniques. For any time series Y_t that is normal (and

therefore linear) $\rho_k(Y_t^2) = \{\rho_k(Y_t)\}^2$ (where $\rho_k(.)$ denotes the lag k autocorrelation). Any departure from this result indicates a degree of non-linearity.

In this data mining study, the influence of soluble total phosphorus (STP) concentration, alkalinity, stream flow (SF), seasonality, man-made intervention (X1), pH, temperature (Temp), dissolved organic carbon (DOC), and dissolved oxygen (DO) at the time of monitoring on the total phosphorus concentration (TP) at six sites (McCarran Bride (MC), Wordsworth Bridge (WB), Derby Dam (DD), Steamboat Creek (SC), Lockwood(LW), and North Truckee Drain (NTD)) in the Truckee River, Nevada was modeled through linear and non-linear mixed models using MIXED, and NLMIXED procedures in SAS® respectively. Data were corrected for missing values, and tested for non-stationarity using the MI procedure and the Augmented Dicky and Fuller test with the ARIMA procedure respectively in SAS® before fitting the linear and non-linear mixed models. Results from linear and non-linear mixed models were compared. Data were tested for normality through the standard tests provided with PROC UNIVARIATE. The developed model can provide a guide to probable range and type of TP load generated and deposited into the Truckee River.

STUDY SITE

The Truckee River can be best described as a river in northern California and northern Nevada, that is140 mi (225 km) long, originates from the mountains, south of Lake Tahoe, flows into the Lake Tahoe at its south end, drains part of the high Sierra Nevada, and empties into Pyramid Lake in the Great Basin (USEPA, 1991). The river passes through the Reno-Sparks metropolitan area, located in Nevada's Truckee Meadows. It flows generally northwest through the mountains to Truckee, California, and then turns sharply to the east and flows into Nevada, past Reno and Sparks and along the northern end of the Carson Range. East of the Truckee Meadows, fourteen ditches remove water for irrigation. The most significant diversion is Derby Dam, where at least 32% of the river's water is diverted annually (Peternel and Laurel, 2005).

Truckee River's waters are an important source of drinking and irrigation along its valley and adjacent valleys. Increased urbanization and the prevalence of water diversions have caused a decline in water quality, and the resulting detrimental effects on habitat have brought about the need to restore the river to a more natural condition to improve habitat and the river's overall health.The water is quite clear near Lake Tahoe, but as it descends, the water turns muddy and concentrated in nutrients and other toxic elements by the time it passes Reno, Nevada. The California State Water Resources Control Board (State board) has classified under Section 330(d) of the Clean Water Act the middle reach of the Truckee River as "impaired" (Dana and others, 2006). Because of the endangered species present and due to the fact that Lake Tahoe Basin comprises the headwaters of the Truckee River, the river has been the focus of several water quality investigations, the most detailed starting in the mid-1980s. Under the direction of the U.S. Environmental Protection Agency, comprehensive dynamic studies have been undertaken to study the impacts of a variety of land use and wastewater management decisions throughout the 3120 square mile Truckee River Basin and also to provide guidance to other U.S. river basins (USEPA, 1991). Analytes mostly addressed include nitrogen, phosphate, dissolved oxygen, and total dissolved solids. Impacts upon, the receiving waters of Pyramid Lake has also been analyzed (Source: Truckee River Geographic Response Plan, 2005).

TMWRF currently maintains 11 continuous monitoring stations within the Truckee water system. These stations are located at: Mogul, Steamboat Creek, McCarran Bridge, North Truckee Drain, Lockwood, Patrick, Waltham, Tracy, Painted Rock, Wadsworth and Marble Bluff Dam. Lockwood monitoring site is located in the lower Truckee River basin 65.6 river miles from Lake Tahoe, located down stream of McCarran Bridge, North Truckee Drain, and Steamboat Creek monitoring sites and Vista (www.tmwrf.com).

WATER QUALITY IN TRUCKEE RIVER

From the results for Truckee River, the total phosphorus concentration (TP), as classified by the Environmental Protection Agency (NDEP, 1994) is a -conservative pollutant. Conservative pollutants are those which persist in the water segment of the aquatic environment over time remaining essentially constant in concentrations. These pollutants are not perturbed by seasonal variations or other short term cyclical and non-cyclical variations in the system. Hence simple techniques are often adequate to set water quality standards for the conservative pollutants (i.e: Total Maximum Daily Loads (TMDL)) in a body of water. For example the concentration based upon an applicable water quality standard can serve as TMDL for a conservative pollutant. In addition the concentration of the conservative pollutant varies directly with the volumes of flows of dischargers of the receiving water body. The TMDL compliance level for total phosphorus concentration for Truckee River is currently at 0.05 mg/L (214 lb/day) at Lockwood monitoring site. Existing data indicate that approximately 80 lb/day are attributable to non-point sources and background. The remaining 134 lb/day were set as the total phosphorus waste-load from the Truckee Meadows Waste Water Reclamation Facility (TMWRF).

LINEAR MIXED EFFECTS MODELS

Mixed linear models provide the flexibility of modeling variances and covariance of variables in addition to means specified in a cross sectional regression model hence can be used to model data that show correlation and non-constant variability. Random effects parameters with non constant variability such as that shown with unbalanced time series cross sectional data (i.e., spatial repeated measures time series data, nested or clustered time series data) can be modeled easily and accurately with PROC MIXED in SAS® which also provides a variety of covariance structures to model random-effects parameters with non constant variability. The most common of these structures arises from the use of random-effects parameters, which are additional unknown random variables assumed to impact the variability of the cross sectional data. The variances of the random-effects parameters, commonly known as variance components, become the covariance parameters for the particular structure. Traditionally mixed linear models were used to model a combination of fixed and random effects that led to the name mixed model. With PROC MIXED in addition to the traditional variance components, numerous other covariance structures are available. PROC MIXED also assumes that the data is: a) normally distributed, contain linear means, and follows specified variance and covariance structure from either a fixed or a random effect. The RANDOM statement with PROC MIXED incorporates random effects constituting the vector type specified in the mixed time series cross sectional model and computes restricted maximum likelihood (REML) and the maximum likelihood (ML) estimates of the variance parameters (Searle 1988; Harville 1988; Searle, Casella, and McCulloch 1992). These estimates can be used to draw statistical inferences about the covariance structure. The REPEATED statement is useful with PROC MIXED to specify covariance structures for cross sectional data where the repeated measures are cross sectional in nature. Mixed models also allow interaction effects of the independent variables in addition to the main effects.

NON-LINEAR MIXED EFFECTS MODELS

Non linear mixed model in addition to the above mentioned features of the mixed models, allow non linear terms in the model and allow conditional distributions of the fixed and the random effects specified. Unlike mixed linear models that require the data be normal non-linear mixed models allow data with a variety of distributions. The NLMIXED procedure in SAS® fits non-linear mixed models by maximizing an approximation to the likelihood integrated over the random effects. The default dual quasi-Newton algorithm was used to fit the non-linear model in this study. Successful convergence of the optimization problem resulted in parameter estimates along with their approximate standard errors based on the second derivative matrix of the likelihood function. The MODEL statement with NLMIXED procedure was used to specify the model terms with the conditional distribution (normal in this case) of the data with the random effects specified. MODEL requires a single dependent variable from the input data set, and a distribution along with its parameters. The BY statement with NLMIXED can be used to perform separate analyses by cross sections identified when the variable site (cross sectional variable) is specified with the BY statement. The CONTRAST statement was used to conduct statistical tests of hypotheses on individual cross sections, with several expressions simultaneously equal zero. The ESTIMATE statement enables to compute an additional estimate that is a function of the parameter values. PROC NLMIXED also computes approximate standard errors for the estimates using the delta method (Billingsley 1986), and the corresponding *t* statistics, *p*-values, and confidence limits. The PREDICT statement was used with PROC NLMIXED to construct predictions of expressions across all of the observations in the input data set. Predicted values are computed using the parameter estimates and empirical Bayes estimates of the random effects. Standard errors of prediction are computed using the delta method (Billingsley 1986; Cox 1998). RANDOM statement was used to define the random effects and their distribution. Random effects typically influence the mean value of the distribution specified in the MODEL statement. RANDOM statement is a list of the random effects, a distribution for the random effects, and a SUBJECT= variable. PROC NLMIXED assumes that a new realization occurs whenever the SUBJECT= variable changes from the previous observation, which requires the input data set be clustered according to this variable. If the input data set is not clustered according to the SUBJECT= variable, data must be sorted using PROC SORT by the SUBJECT= variable prior to calling PROC NLMIXED. Since only one RANDOM statement is permitted with PROC NLMIXED multilevel nonlinear mixed models are not allowed. In addition there is only one distribution (normal(m,v) [m=mean and v=variance]) currently allowed for the random effects with the RANDOM statement. However, multiple effects can be specified as bracketed vectors for means and variances.

NON-LINEAR OPTIMIZATION

A commonly encountered problem with NLMIXED is that the estimated covariance matrix of the parameter estimates is computed as an inverse Hessian matrix, which needs to be positive definite for unconstrained problems. SAS® allows three singularity criteria (ASINGULAR=, MSINGULAR=, and VSINGULAR= options) to fine tune the inversion of the Hessian matrix. The rows and columns of the Hessian matrix can also be scaled while using the optimization techniques, with the HESCAL= option. There are several optimization techniques (Newton- Raphson, Quasi-

Newton, double dog-leg, conjugate gradient, Nelder-Mead Simplex) available with PROC NLMIXED. A particular optimization method can be selected with the TECH= option in the PROC NLMIXED statement. The default Quasi Newton (QUANEW) was used in the present study with the DBFGS (dual Broyden, Fletcher, Goldfarb, and Shanno (BFGS) update of the Cholesky factor of the Hessian matrix) update method. QUANEW provides an appropriate balance between the speed and stability required for most nonlinear mixed model applications. Nonlinear optimization can be computationally expensive in terms of time and memory, so the appropriate optimization algorithm must be carefully matched to the problem. NRRIDG algorithm is the fastest among the second derivative methods (TRUREG, NEWRAP, NRRIDG) and also requires only one matrix with n(n+1)/2 double words (requires less memory) hence the best for small problems. For small problems the Hessian matrix is not expensive to compute. The first-derivative methods QUANEW and DBLDOG are best for medium-sized problems where the objective function and the gradient are much faster to evaluate than the Hessian. The QUANEW and DBLDOG algorithms, in general, require more iterations than TRUREG, NRRIDG, and NEWRAP, but each iteration can be much faster. The QUANEW and DBLDOG algorithms require only the gradient to update an approximate Hessian, and they require slightly less memory than TRUREG or NEWRAP (essentially one matrix with n(n+1)/2 double words). The default optimization method is QUANEW. The first-derivative method CONGRA is best for large problems where the objective function and the gradient can be computed much faster than the Hessian and where too much memory is required to store the (approximate) Hessian. The CONGRA algorithm, in general, requires more iterations than QUANEW or DBLDOG, but each, iteration can be much faster, and requires only a factor of n double-word memory. Hence many large applications using PROC NLMIXED can be solved only by CONGRA.

OBJECTIVES

1) To identify and model non-linearity in the total phosphorus concentration time series collected over six selected monitoring sites at the Truckee River in Nevada, using non-linear time series analysis.
2) To predict the influence of soluble total phosphorus, alkalinity, dissolved organic carbon, dissolved oxygen, pH, temperature, seasonality, intervention, and stream flow, on the total phosphorus concentration in Truckee River, Nevada.

EXPLORATORY DATA ANALYSIS

ORIGINAL DATA

Monthly values of TP, STP, DOC, and DO in milligrams per liter (mg/L), and Alkalinity, pH, and Temp, and SF in cubic feet per second (cfs) for the period from January 1997 through December 2004 were obtained from the Truckee River Water Reclamation Facility for the Truckee River at six selected sampling sites namely: i) McCarran bridge (MC), ii) North Truckee Drain (NTD), iii) Wordsworth bridge (WB), iv) Lockwood (LW), v) Derby Dam (DD), and vi) Steamboat Creek (SC) as EXCEL spread sheets (www.tmwrf.com) and converted to SAS® data sets.

ORIGINAL SERIES PLOTS

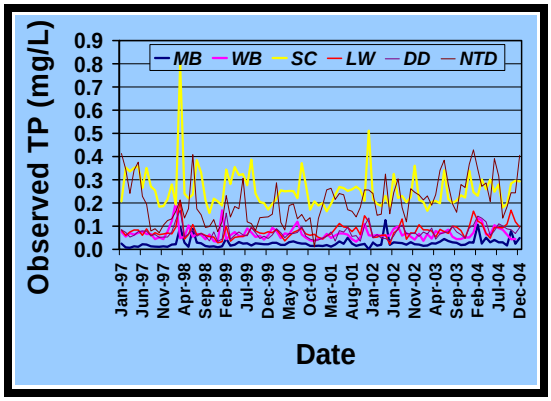

Figure 1: Original TP in between January 1997 and December 2004 by site

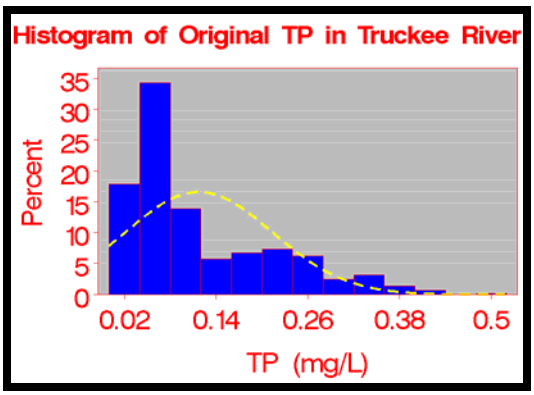

Figure 2: Histogram of original TP at all sites

There are significant fluctuations in TP among the sites (Figure 1). Distribution of original TP is not normal curvilinear (Figure 2). Box plot of the distribution of original observed TP in Truckee River by site is shown in Figure 3. The one observation (observation 205) that had a TP concentration (0.799) above 0.75 was considered an outlier

and was removed from the dataset. Box plot of the distribution of original TP by site after removing the outlier is shown in Figure 4. Mean and maximum TP at site SC remains much above the overall mean (0.117 mg/L; Table 1) and overall maximum (0.512 mg/L; Table 1). Maximum TP values at sites DD, LW, MB, NTD and WB except at site SC are below 0.5 mg/L. TP values above 0.3 mg/L were considered extreme and an intervention analysis was considered. Site SC shows the largest TP during most part of the time period studied, followed by NTD. Past studies have indicated that high phosphorus loads, associated largely with TMWRF and irrigation return flows, to significantly impact the Truckee River. Figures 5 and 6 present monthly average TP and DO concentrations at LW and SC monitoring sites respectively. These elevated phosphors loads along with elevated nitrogen loads can encourage proliferation of aquatic plants and benthic algae. Respiration by these plants and the decay of their associated detritus can decrease DO in the water column, resulting in violations of the DO standard (5mg/L). The mean, maximum and minimum values of TP at LW monitoring site are 0.078,, 0.212, 0.022 mg/L respectively and that at SC monitoring site are 0.257, 0.799, 0.157 mg/L respectively. The mean, maximum, and minimum values of DO at SC monitoring site are 8.83, 13.2, and 4.5 mg/L respectively and that at LW monitoring site are 10.59, 13.6 and, 7.8 mg/L respectively.

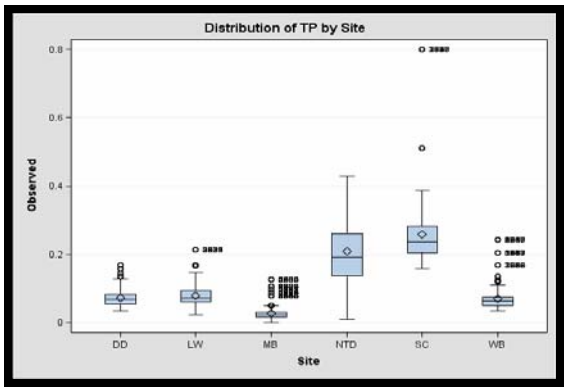

Figure 3: Box plot of distribution of observed TP by site

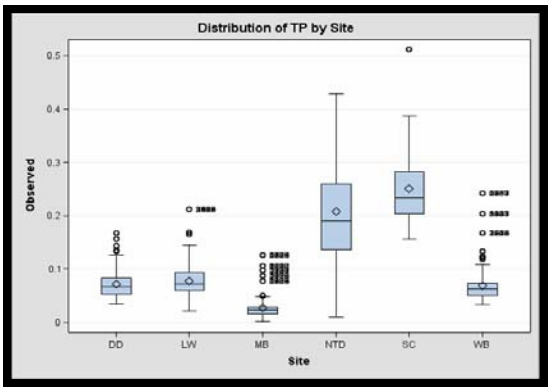

Figure 4: Box plot of distribution of outlier removed TP by site

The TP concentrations have remained above the compliance level at SC throughout the study period. Mean DO concentrations are also above 5 mg/L at both monitoring sites. DO levels are above 5 mg/L at LW throughout the study period. This effect may be due to low total nitrogen concentrations in the River during the study period. Low DO has caused several fish kills, impacting populations of Lahontan cutthroat trout and cui-ui (kwee-wee) in the past (Adele Basham, NDEP, personal communication, 1994). The cui-ui, a fish that has historically been a staple in the diet of the local Pyramid Lake Paiute Tribe, is on the national endangered species list. The Lahontan cutthroat trout, which supports a small recreational fishery, is classified as a threatened species.

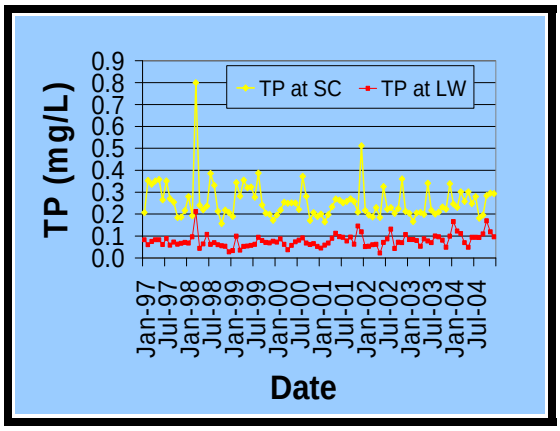

Figure 5: Original TP between January 1997 and December 2004 at SC and LW

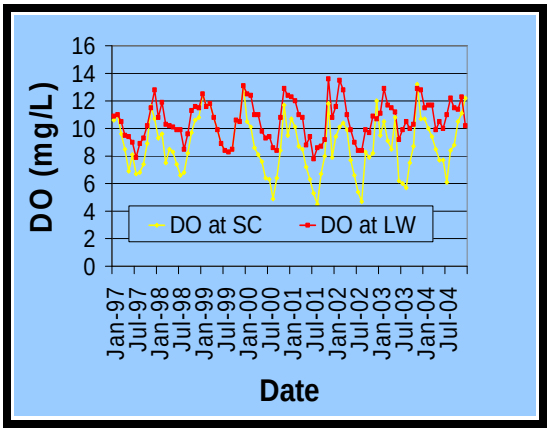

Figure 6: Original DO between January 1997 and December 2004 at SC and LW

SE

Presence of unit roots (or non-stationarity) has critical consequences in time series data analysis. Estimating a model with either a lagged dependent variable or serially correlated errors in the presence of a unit root can lead to dramatically misleading results (spurious regressions). Checking if the residuals appear stationary (that is, whether an autoregression of the residuals on their lags shows a coefficient on the lagged residual term near one), as well as examine whether the coefficients on any lagged dependent variable terms are near one is essential in any time series data analysis. The augmented Dicky and Fuller test was performed to test the data for the presence of unit root non-stationarity

According to (Augmented) Dicky and Fuller test a pure random process $X_t = \rho X_{t-1} + \mu_t$, is non-stationary whenever $\rho = 1$ which is called the existence of unit root. In other words the process can be described as: $X_t - X_{t-1} = (\rho-1) X_{t-1} + \mu_t$, and also $\Delta X_t = \delta X_{t-1} + \mu_t$. Whenever $\delta=0$, unit root exists and the process is non-stationary. ' t' value of the coefficient of X_{t-1} follows a 'tau' statistic. There are three methods available to correct the data for non-stationarity, a) differencing the data, b) data transformation, and c) using sinusoidal components. Differencing the data is the most common method, which was used to achieve stationarity in this study. By differencing a time series we 'derive' the function by which the original series was generated, which removes any trend in the mean level and any seasonal and non-seasonal cycles and periodicities found in the original series. First differencing converts the above pure random process to $X_t - X_{t-1} = \mu_t$, hence to a stationary process whenever, $\rho=1$. The augmented Dicky and Fuller non-stationarity test (SAS CODE 1) indicated the presence of non-stationary in the TP series at 5% level ($p<0.0001$). First differencing was found adequate to correct data for non-stationarity (SAS CODE 2). First differencing corrected the data for non-normality and non-stationarity. The normal histogram plot of the overall TP series after first differencing shows normality (hence linear) in first differenced series (Figure 7). Significant non-linear autocorrelations existed in the TP series ($p<0.0001$) (Figure 8), that were removed by first differencing the series (Figure 9). Box plot of distribution of TP by site after first differencing is shown in Figure 10.

SAS® CODE 1

```
PROC ARIMA DATA=comp;
      IDENTIFY VAR=TP(1)
      STATIONARITY=(ADF=(1,2,4,6,12))
      CROSSCORR=(Summer X1 STP Alkalinity
       DOC DO SF Temp pH);
      ESTIMATE P=(1)  Q=(1) NOINT
      OUTLIER MAXNUM=5 ALPHA=0.01;
```

SAS® CODE 2

```
DATA comp; SET comp;
   TP=DIF(TP);
 RUN;
```

MISSING VALUES

Missing values are an issue in time series analyses. Most time series models (both linear and nonlinear) require non missing data. Performing analyses with incomplete cases also ignore possible systematic differences between the complete cases and the incomplete cases, and the resulting inference may not be applicable to the population of all cases, especially with a small number of complete cases. PROC MI in SAS® was used to correct the data for missing values using the Markov Chain Monte Carlo (MCMC) method (Schafer 1997) that assumes multivariate normality (SAS CODE 3). MCMC method requires the assumption of an arbitrary missing pattern to impute all missing values.

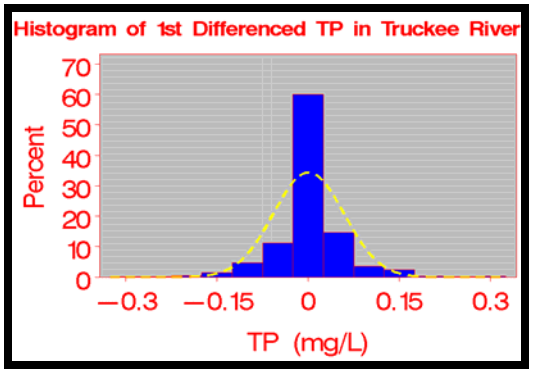

Figure 7: Histogram of the first differenced TP at all sites

SAS® CODE 3

```
PROC MI DATA=comp SEED=21355417
NOINT NIMPUTE=6 MU0=50 10 180
   OUT=outmi;
MCMC CHAIN=multiple DISPLAYINIT
   INITIAL=em(ITPRINT);
 VAR TP Alkalinity DO2 DOC STP
     SF pH Temp;
 RUN;
```

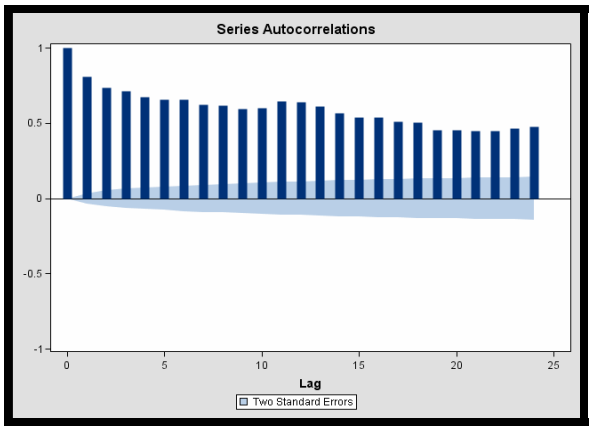

Figure 8: Autocorrelation function plot of non-differenced outlier removed TP series

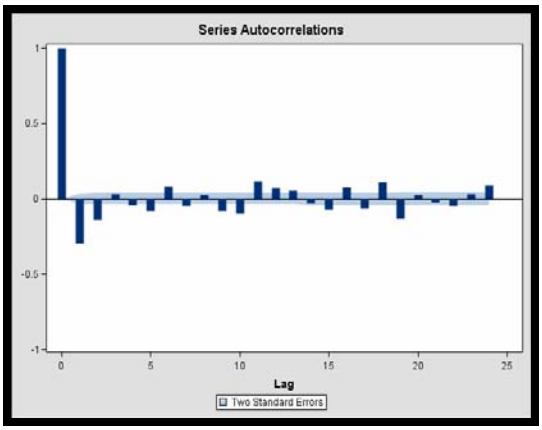

Figure 9: Autocorrelation function plot of first differenced outlier removed TP series

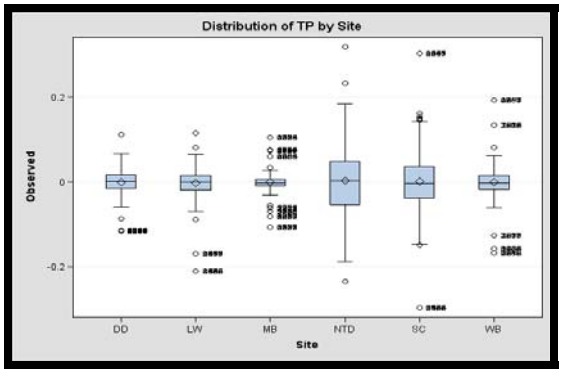

Figure 10: Box plot of distribution of first differenced outlier removed TP

Table 1: Basic statistical measures of outlier removed TP

Location		Variability	
Mean	0.1173	Std Deviation	0.0961
Median	0.0770	Variance	0.0092
Mode	0.0630	Range	0.5110
Maximum	0.5120	Inter-quartile Range	0.1330
Minimum	0.0010	Coeff. Variation	81.880

PROC MI performs multiple imputation of missing data. Which means instead of filling in a single value for all missing points each missing value is replaced with a set of plausible values that represent the uncertainty about the right value to impute.

SAS® CODE 4

```
PROC UNIVARIATE DATA=outmi
                NORNALTEST;
     VAR TP;
     HISTOGRAM TP /
     NORMAL(COLOR=
       yellow w=3)
     CFILL= blue  CFRAME=ligr
     NAME='First differenced
       TP at Truckee River';
     OUTPUT OUT=outuni MEAN=
        MTP STDERR= STDTP;
RUN;
```

Table 2: Goodness-of-Fit Tests for Normal Distribution: First differenced TP

Test	Statistic	p Value
Kolmogorov-Smirnov	0.172157	<0.010
Cramer-von Mises	38.565848	<0.005
Anderson-Darling	195.850732	<0.005

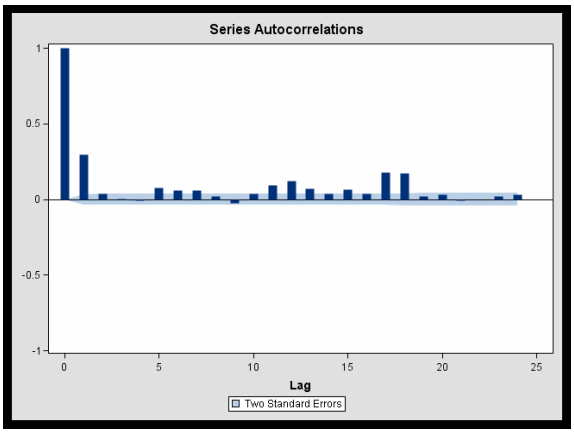

Figure 11: Autocorrelation function plot of the square transformed outlier removed first-differenced TP

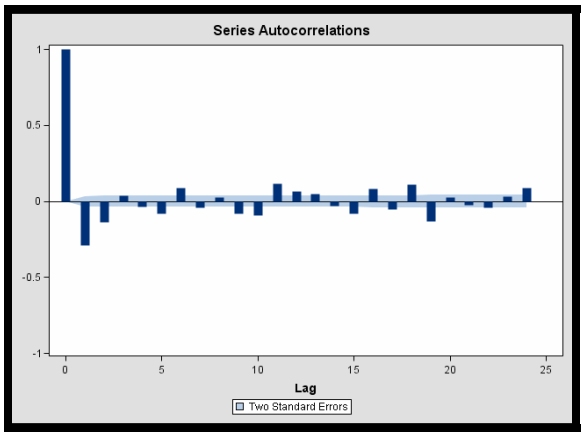

Figure 12: Autocorrelation function plot of exponentially transformed outlier removed first-differenced TP

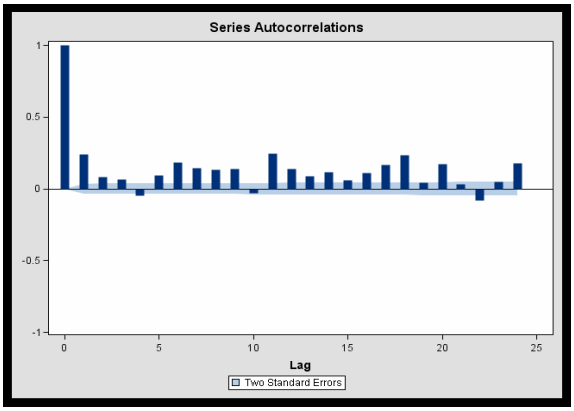

Figure 13: Autocorrelation function plot of log transformed outlier removed first-differenced TP

SAS® CODE 5

```
PROC MIXED DATA=MONTHLY METHOD=
                ML COVTEST;
CLASS SITE DATE;
MODEL TP= DOC DO SF pH STP
    SUMMER X1 TEMP Alkalinity
    DO*SF DOC*SF pH*DOC
    X1*SF Temp*SF pH*SF
    Alkalinity*DO Alkalinity*SF
    Alkalinity*DOC Alkalinity*STP
   /INFLUENCE EFFECT=SITE) S;
REPEATED /TYPE=CS SUBJECT=SITE R;
RUN;
```

Multiply imputed data sets can be analyzed by using any standard procedures for complete data combining the results from the analyses. Multiple imputation, also does not attempt to estimate each missing value through simulated values. Instead, a random sample of missing values, are drawn from its distribution. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, confidence intervals with the correct probability coverage. In addition multiple imputation inference assumes that the model (variables) used to analyze the multiply imputed data is the same as the model used to impute missing values (the imputer's model) which may not be the case actually. Multiple imputation variance information and imputation parameter estimates were obtained for all variables from MI procedure. Results indicated high efficiency of the MCMC method (relative efficiency varied between 0.86 and 0.99935). Relative increase in variance was zero for all variables. Imputation estimates were highly significant at the 5% level (P<0.0001).for all variables.

UNIVARIATE ANALYSIS FOR NORMALITY

A univariate normal analysis was performed using PROC UNIVARIATE with the NORMALTEST) (SAS CODE 4) option in SAS® on the combined (over six sites) outlier removed first differenced TP series (hereafter referred as TP) to obtain the overall basic statistical measures (mean, standard deviation etc.) of the data (Table 1) and to test the variable for normality using standard normality tests (Anderson-Darling, Kolmogorov-Smirnov etc.) provided in SAS® (Table 2). The probability of all test statistics are smaller than 0.05 (Table 2) and the hypothesis of normality of the data was accepted. A histogram of the variable (TP) was also requested using the HISTOGRAM and NORMAL options with a normal curve superimposed on the histogram (Figure 7).

NON LINEARITY

A simple yet powerful way to identify non-linearity in time series data is to plot the auto correlation function (ACF) of the non-linearly transformed (square, exponential, log) normal (hence linear) series. Significant spikes in the ACF plot at specific lags in the non-linearly transformed linear series indicate the presence of non-linearity. ACF plots of the linear first differenced TP series were obtained after non-linearly transforming (squared, exponential, log) the series. ACF plots indicate the presence of non-linearity in the TP series (Figures 11, 12, and, 13), Relationship of one or more of the independent variables to the dependent variable is non-linear. Appropriate non-linear model has to be identified through trial and error.

SEASONALITY AND MAN-MADE INTERVENTION MODELING

Influence of summer (summer) and winter (winter) seasonality and man-made intervention (X1) on TP concentration were considered for inclusion in the linear and non-linear mixed effects models as explanatory variables along with other independent variables. TP values above six times the compliance level (0.3) were considered extreme and a man-made intervention on these observations was calculated. X1, summer and winter variables were computed as shown below for the mixed effects models as explanatory variables. Only the summer seasonality variable was used in the final model since the influence of the winter on TP was the negative of the influence of summer on TP.

X1 = 'man made intervention' ; summer = 'summer months'; winter = 'winter months'
X1 = TP >0.3; summer = (5 < mm < 11) * (year > 1995); winter = (year > 1995) - summer;

LINEAR MIXED MODEL APPLICATION

A linear mixed model was fitted to the combined time series data to analyze the overall and cross sectional influence of the independent variables on TP in Truckee River. SAS® procedure MIXED was used with the REPEATED option to analyze the effect of all independent variables. The REPEATED statement assumes that the observations are ordered similarly for each site. The TYPE=CS option specifies a compound symmetry (CS) covariance structure for each SUBJECT=Site (SAS CODE 5). The CS covariance structure was tested (likelihood ratio tests) against and found superior to Unstructured (UN), first order Autoregressive (AR(1)), and Exponential error structures. The CS structure has two covariance parameters and a common variance with a diagonal enhancement that results from 50*50 compound symmetric structured blocks. The overall mixed model contains 6 cross sectional sites and 96 time series data values within sites. The **R** matrix (variance/covariance matrix) is, therefore, block diagonal with 6 blocks, each block consisting of identical 50×50 cross sectional matrices and the data are balanced. The R option with REPEATED requests that the first block of **R** be displayed (SAS CODE 5).

Model convergence was achieved in 4 iterations. Overall model was significant. Individual cross sectional effects of the six sites are shown in Table 3. All sites contribute significantly at 5% level to overall TP concentration in Truckee River. Model fit statistics show adequate statistical fit of the model (Table 4). Overall influence of DOC, DO, SF, STP, X1, and Temp on TP concentration are highly significant at the 5% level ($p<0.05$, Table 5). The two way interaction effects among all the independent variables studied were highly significant at the 5% level ($p<0.05$). Seasonality (summer), pH and DO are negatively correlated to TP (Table 4). All other independent variables are positively correlated to TP. DO is negatively correlated to TP concentration as expected. The relationship between pH and TP, alkalinity and TP and seasonality and TP could not be predicted by the linear mixed effects model. These relationships may be non-linear. Scatter plots of TP versus pH and TP versus alkalinity of observed outlier removed data were obtained for further analysis for non-linear model fitting (Figures 14 and 15). TP does not change with pH. Larger TP values are associated with larger alkalinity values. The positive linear relationship prevalent between TP and alkalinity cannot be predicted by the linear mixed effects model. The effect of seasonality also could not be predicted by the linear mixed model.

NON-LINEAR MIXED MODEL APPLICATION

Non-linear mixed effects model was fitted to the data using PROC NLMIXED. According to the non-linear mixed model (SAS® CODE 6) larger number of independent variables significantly (at 1% level) influence TP in Truckee River (Table 6) than predicted by the linear mixed model (Table 5). Larger percent of the variation in the data was explained by the non-linear fixed effects compared to the linear fixed effects at 1% significance level ($p<0.0001$). According to the non-linear mixed model pH ($p<0.0001$) and alkalinity (ALK.) ($p<0.0001$) significantly influence TP in Truckee River at 5% level. Results indicate that non-linear effects could be important in time series regression and modeling. Parameter estimates of all the studied independent variables are highly significant at 1% level (Table 6). Much smaller gradients that are significant at the 1% level were detected (Table 6). Model diagnostic

statistics are much smaller compared to that calculated for the linear mixed effects model (Table 7). Model convergence was achieved in 73 iteration steps with the selected optimization technique.

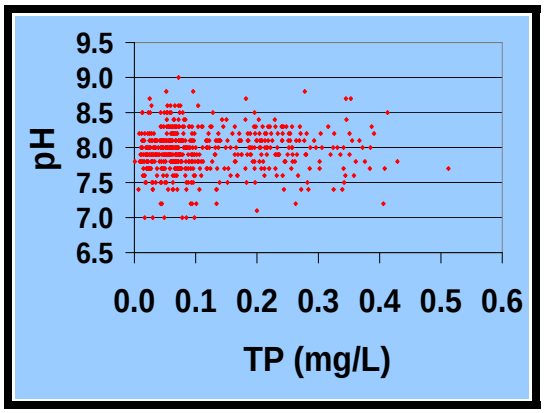

Figure 14: Scatter plot of TP versus pH of outlier removed first-differenced TP

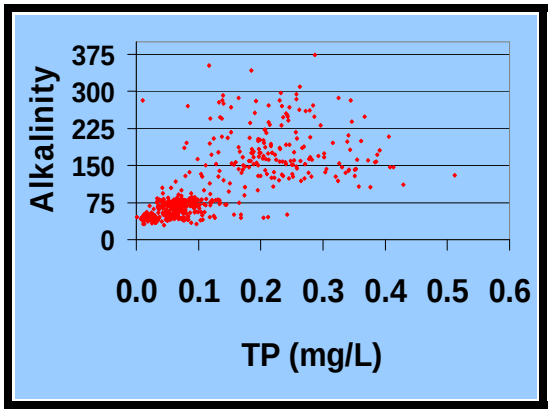

Figure 15: Scatter plot of TP versus alkalinity of outlier removed first-differenced TP

Table 3: Individual Site effects (Linear Mixed effects model)

Site	Estimate	Std Error	t Value	Pr > |t|
MB	0.4846	0.08581	5.65	<.0001
WB	0.4944	0.07110	6.95	<.0001
SC	0.5348	0.06996	7.64	<.0001
LW	0.1529	0.04243	3.60	0.0003
DD	0.1006	0.03604	2.79	0.0053
NTD	0	-	-	-

Table 4: Fit Statistics (Linear Mixed Model)

-2 Log Likelihood	-16517.1
AIC (smaller is better)	-16481.1
AICC (smaller is better)	-16480.9
BIC (smaller is better)	-16484.8

Model predicts observed TP fairly accurately (Figure 16). **The** Contribution from individual cross sectional sites to overall TP in Truckee River is highly significant ($p<0.0001$) at 1% level of significance (Table 8). All six sites contribute positively to overall TP concentration in Truckee River. DO at SC and LW are negatively significantly ($p<0.0001$) correlated to TP in Truckee River at 1% level (Table 8). SF from NTD positively significantly ($p<0.0001$) influences TP in Truckee River.

FITTED NON-LINEAR MIXED MODEL

The following non-linear model was found adequate and superior to linear mixed models to predict TP in Truckee River.

```
SF=log(abs(SF));
Summer=exp(Summer-12);
X1=exp(X1-12);
Alkalinity=(1-Alkalinity);
Temp=(1/(1-Temp));
DOC=log(abs(1/DOC));
DO=exp(DO);
```

$$TP = beta1 + beta2*DOC + beta3*DO + beta4*SF + beta5*pH + beta6*STP + beta7*Summer + beta8*X1 + beta9*Temp + beta10*Alkalinity + beta11*DO*SF + beta12*DOC + beta13*doc*ph + beta14*pH*SF + beta15*X1*SF + beta16*Temp*SF + beta17*DO*STP + beta18*STP*SF + beta19*Alkalinity*DO2 + beta20*Alkalinity*SF + beta21*Alkalinity*DOC + beta22*Alkalinity*STP \quad [1]$$

The description of the coefficients (betas) in the above model (Eq. 1) and their estimates are given in Tables 6.

Table 5: Parameter Estimates (Linear Mixed Model)

Effect	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.000246	0.000223	1.10	0.3204
Alkalinity	-0.00002	0.000017	-0.89	0.3755
DOC	0.006199	0.000296	20.92	<.0001
DO	-0.00103	0.000512	-2.01	0.0444
SF	8.83E-6	0	Infty	<.0001
pH	-0.00038	0.001301	-0.29	0.7700
STP	0.7867	0.01558	50.50	<.0001
Summer	-0.00160	0.001295	-1.24	0.2167
x1	0.03665	0.002127	17.23	<.0001
Temp	0.000618	0.000175	3.53	0.0004
DO*SF	3.854E-6	0	Infty	<.0001
DO*STP	-0.01493	0.007538	-1.98	0.0477
DOC*SF	-8.98E-8	0	-Infty	<.0001
SF*STP	0.000105	0.000019	5.67	<.0001
Alkalinity*DO	-0.00004	0.000012	-3.17	0.0015
Alkalinity*SF	4.743E-9	0	Infty	<.0001
Alkalinity*DOC	0.000015	4.645E-6	3.18	0.0015
Alkalinity*STP	0.000965	0.000197	4.90	<.0001
DOC*pH	-0.00337	0.000592	-5.68	<.0001
SF*pH	-8.57E-6	2.142E-6	-4.00	<.0001
SF*X1	-0.00002	2.904E-6	-7.33	<.0001
SF*Temp	3.042E-6	0	Infty	<.0001

Table 6: Parameter Estimates (Non-linear mixed model)

Parameter	Variable	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Gradient
beta1	Intercept	-0.5415	0.05013	560	21.57	<.0001	0.01	-0.057
beta2	DOC	-0.1582	0.006053	560	-2.72	<.0001	0.01	0.092024
beta3	DO	-2.22E-8	0.001628	560	3.70	<.0001	0.01	2171.363
beta4	SF	0.08010	0.002206	560	124.84	<.0001	0.01	-0.35136
beta5	pH	0.06026	0.006022	560	-2.89	<.0001	0.01	-0.46-97
beta6	STP	1.4832	0.000999	560	-521.47	<.0001	0.01	-0.00786
beta7	Summer	-0.1832	0.003326	560	1.36	<.0001	0.01	0.013948
beta8	X1	0.3180	0.006041	560	-10.47	<.0001	0.01	-0.01384
beta9	Temp	-0.03603	0.000387	560	3.46	<.0001	0.01	0.008352
beta10	Alkalinity	-0.00044	0.000129	560	-7.37	<.0001	0.01	11.84719
beta11	DO*SF	5.967E-9	1.857E-6	560	-3.90	<.0001	0.01	13254.31
beta12	DOC*SF	-0.00062	0.006053	560	-3.74	<.0001	0.01	0.578847
beta13	DOC*pH	0.01988	0.001525	560	2.68	<.0001	0.01	0.742752
beta14	pH*SF	-0.00872	2.801E-6	560	3.24	<.0001	0.01	-2.79159
beta15	X1*SF	1.6995	0.000367	560	114.90	<.0001	0.01	-0.08361
beta16	Temp*SF	0.005920	0	560	Infty	<.0001	0.01	0.040595
beta17	DP*STP	9.156E-7	0.007051	560	-12.21	<.0001	0.01	149.9421
beta18	STP*SF	-0.00385	0.000015	560	3.81	<.0001	0.01	-0.05008
beta19	Alk.*DO	6.87E-10	0.000011	560	5.60	<.0001	0.01	-358508
beta20	Alk. *SF	0.000061	0.000365	560	-8.62	<.0001	0.01	72.59328
beta21	Alk.*DOC	0.000124	0.000010	560	1.81	<.0001	0.01	-19.084
beta22	Alk.*STP	0.002068	0.000345	560	5.92	<.0001	0.01	1.657662

SAS® CODE 6

```
PROC NLMIXED DATA=MONTHLY QPOINTS=10 ALPHA=0.05 TECH=QUANEW;
 BY SITE;
   PARMS
      beta1= 0.000246 beta2= 0.006199 beta3 = -0.000000548
      beta4= 0.0000000 beta5=-0.0000002455 beta6= -0.7867
      beta7= 0.1 beta8= 0.03665 beta9= -0.00066 +
      beta10=-0.002 beta11=0.000003854 beta12=0.0000000898
      beta13=-0.00337 beta14=-0.00000857 beta15=0.00002
      beta16= 0.000003042 beta17=0.01493 beta18=0.000105
      beta19= 0.00004 beta20=0.000000004743 beta21=0.000015
      beta22= 0.000965 g11=-0.001428 to 0.02 by 0.001
      g12=-0.001 to 0.01 by 0.001;

        SF=log(abs(SF));
        Summer=exp(Summer-12);
        X1=exp(X1-12);
        Alkalinity=(1-Alkalinity);
        Temp=(1/(1-Temp));
        DOC=log(abs(1/DOC));
        DO=exp(DO);

  eta=  beta1+ beta2*DOC + beta3*DO + beta4*(SF) +
        beta5*pH + beta6* STP + beta7*(Summer) + beta8*X1 +
        beta9*Temp + beta10* Alkalinity + beta11*DO*SF +
        beta12*DOC + beta13*DOC*pH + beta14*pH*SF +
        beta15*DO*SF + beta16*Temp*SF + beta17*DO*STP +
        beta18*STP*SF + beta19*Alkalinity*DO +
        beta20*Alkalinity*SF + beta21*Alkalinity*DOC +
        beta22*Alkalinity*STP; /* +b1; */
  num= eta;
  mu=  num;

    MODEL TP ~ NORMAL(mu,g12);
  /* RANDOM b1 ~ NORMAL(0,g11) SUBJECT=SITE;*/
    PREDICT mu OUT= predmean;
    CONTRAST 'LW vs MB' beta26-beta23;
 /*ODS SELECT CorrMaTParmEst CovMaTParmEst;*/
RUN;
```

Results of mean variable contrasts indicate that mean TP at LW site is significantly larger ($p<0.05$) than mean TP at MB and WB, and significantly smaller ($p<0.05$) than the mean TP at SC, and NTD at 5% level of significance (Table 9). Mean TP at LW, SC, NTD, and DD are also significantly different from the overall mean TP in Truckee River. Observed and model predicted mean TP values at each site are shown in Table 10 for comparison. Predicted mean TP at SC and NTD are significantly larger ($p<0.0001$) than overall mean TP and significantly smaller ($p<0.0001$) than overall mean TP.

Table 7: Fit Statistics (Non-Linear MIXED Model)

-2 Log Likelihood	-2493
AIC (smaller is better)	-2445
AICC (smaller is better)	-2442
BIC (smaller is better)	-2341

Table 8: Parameter Estimates of individual sites (Non-linear mixed model)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Alpha	Gradient
MB	0.3096	0.01570	19.71	<.0001	0.01	0.000143
WB	0.5027	0.01069	47.02	<.0001	0.01	-0.0003
SC	0.5731	0.006094	94.04	<.0001	0.01	0.0003
LW	0.2150	0.01141	18.84	<.0001	0.01	-0.0002
DD	0.1672	0.01265	13.22	<.0001	0.01	0.000187
NTD	0.000100	0	Infty	<.0001	0.01	-0.00013
SC*DO	-0.06550	0.009074	-7.22	<.0001	0.01	-0.00041
LW*DO	-0.1796	0.01604	-11.19	<.0001	0.01	0.000174
NTD*SF	0.000085	4.481E-6	18.89	<.0001	0.01	-0.75922

Table 9: Differences in mean TP (Individual sites and overall): Non-linear mixed model

Description (mean TP at)	Num DF	Den DF	F Value	Pr > F
LW vs MB	1	575	33.67	<.0001
LW vs WB	1	575	10.33	0.0014
LW vs SC	1	575	11.88	0.0006
LW vs DD	1	575	2.96	0.0859
LW vs NTD	1	575	8.86	0.0030
LW vs overall TP	1	575	17.60	<.0001
SC vs overall TP	1	575	45.78	<.0001
NTD vs overall TP	1	575	7.83	0.0053
DD vs overall TP	1	575	16.19	<.0001

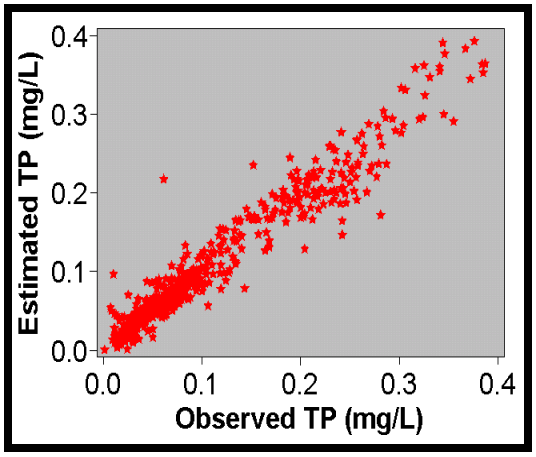

Figure 16: Observed versus model predicted TP in Truckee River (non-linear mixed model)

Table 10: Predicted and observed mean TP at Truckee River

Site	Predicted mean TP (mg/L)	Observed mean TP (mg/L)
MB	0.029	0.027
WB	0.067	0.069
SC	0.228	0.248
LW	0.085	0.078
DD	0.072	0.072
NTD	0.218	0.208
OVERALL	**0.114**	**0.117**

CONCLUSIONS

Non-linearity in fixed and random effects in continuous time series data can affect the conclusions drawn from the results of time series regression. Cross sectional variation in water quality is a key factor in non-point source pollution control studies. Time series data need be tested and corrected for non-linear autocorrelations before performing a regression analysis. An alternative approach is to fit a non-linear model to the observed time series which was the approach used in this study. PROC NLMIXED was used to include non-linear terms in mixed effects models. Results from the non-linear mixed model were compared to that obtained from the linear mixed effects model. According to the results the non-linear modeling approach is much superior to the linear mixed effects modeling approach. Much larger portion of the total variation was explained by the non-linear mixed effects model. NLMIXED procedure is a powerful technique for fitting non-linear mixed models to continuous time series with multiple independent variables.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification, IEEE trans.: Autom. Control, AC-19, 716-723.
Billingsley, P. (1986), Probability and Measure, Second Edition, New York: John Wiley & Sons, Inc
Box, G.E.P., and Jenkins, G.M. (1976) Time series Analysis Forecasting and Control, (2nd ed.): Holden-

Day, San Francisco, Ca.
Buse, A. (1973) Goodness of Fit in Generalized Least Squares Estimation, American Statistician, 27, 106-108.
Cox, C. (1998), "Delta Method," Encyclopedia of Biostatistics, Eds. Peter Armitage and Theodore Colton, New York: John Wiley, 1125-1127.
DaSilva, J.G.C. (1975) The Analysis of Cross-Sectional Time Series Data, Ph.D. dissertation, Department of Statistics, North Carolina State University.
Fuller, W. (1978) Introduction to Time Series, New York: John Wiley & Sons, Inc.
Harville, D.A. (1988) Mixed-Model Methodology: Theoretical Justifications and Future Directions, Proceedings of the Statistical Computing Section, American Statistical Association, New Orleans, 41-49.
NDEP (1993a). Truckee River strategy. Nevada Division of Environmental Protection, Carson City, Nevada.
NDEP (1993b). Truckee River final total maximum daily loads and waste load allocations. Nevada Division of Environmental Protection, Carson City, Nevada.
NDEP (1994) Truckee River final total maximum daily loads and waste load allocations. Nevada Division of Environmental Protection, Carson City, Nevada.
Nowlin, J.O. (1987) Modeling nutrient and dissolved oxygen transport in the Truckee River and Canal downstream from Reno, Nevada. U.S. Geological Survey Water Resources Investigations Report 87-4037. Carson City, Nevada.
Parks, R.W. (1967) Efficient Estimation of a System of Regression Equations When Disturbances Are Both Serially and Contemporaneously Correlated, Journal of the American Statistical Association, 62, 500-509.
Peternel, K., and Laurel, S. (May 15-May 19, 2005) Truckee River Restoration Modeling, World Water and Environmental Resources Congress. Anchorage, Alaska, USA.
Schafer, J.L. (1999) Multiple Imputation: A Primer, Statistical Methods in Medical Research, 8, 3-15.
Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
Searle, S. R. (1988) Mixed Models and Unbalanced Data: Wherefrom, Whereat, and Whereto?, Communications in Statistics - Theory and Methods, 17(4), 935-968.
Searle, S. R., Casella, G., and McCulloch, C.E. (1992) Variance Components, New York: John Wiley & Sons, Inc.
Truckee Meadows Water Reclamation Facility: www.tmwrf.com
Truckee River Geographic Response Plan, 2005: http://ndep.nv.gov/bca/emergency/truckee_river_plan05.pdf
USEPA. (1989) Regionalization as a tool for managing environmental resources. EPA 600/3-89-060. U.S. Environmental Protection Agency, Washington, DC.
USEPA (1991) Guidance for water quality-based decisions: The TMDL process. EPA 440/4-91-001. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Name: Anpalaki J. Ragavan M.S.
Enterprise: Department of Mathematics and Statistics, University of Nevada
Address: 3925, Clear Acre Lane, # 188,
Reno, NV 89512, USA.
Work phone: (775)-327-5260
Home phone: (775)-674-0397
Email: ragavan@unr.edu
Web: None