Paper 205-31

# Modeling Water Quality Trend in Long Term Time Series

Anpalaki J. Ragavan, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557,
George C. Fernandez, College of Agriculture, Biotechnology and Natural Resources, University of Nevada, Reno, NV 89557

## ABSTRACT

Water quality variables change continually through time, arise from dynamic processes and consist of random error components with stochastic variations in space and time that cannot be modeled or explained by normal analytical procedures. Water quality time series with long-term trend, when recorded by any consistent time interval, will display some measure of auto-correlation. This is expected to affect the p-values derived from autoregressive and/or the moving average time series model parameters.

Parametric seasonally adjusted ARIMA model was successfully applied to predict the long-term trend of water quality of a randomly selected river using the SAS®/ETS software. The EXPAND procedure in SAS® was used to adjust the data for missing values. The PROC ARIMA in SAS® was used to correct the data for non-stationarity, and to identify model orders, and the type of seasonal and non-seasonal differencing required. Data were seasonality adjusted through seasonal differencing. An ARIMA model was fitted to the trend removed and seasonally adjusted series using the MODEL procedure in SAS® using the maximum likelihood (ML) method. The use of SAS® procedure UCM for stochastic time series modeling was introduced.

Model parameters were significant with non-significant correlations at 5% level. Goodness of fit statistics indicated adequate statistical fit of the model. No autocorrelations were detected in the residuals estimated from the model. Residuals were normal, and white noise. Time series modeling with ARIMA models provides a powerful option for long-term water quality trend using the SAS®/ETS software.

## INTRODUCTION

Long-term trend of water quality in natural systems reveal information about chemical and biological changes and variations due to man made and/or seasonal interventions. The success of such trend analysis depends largely on the initial exploratory analysis of the data and in identifying the appropriate model orders to predict the trend [5]. The best approach currently available to model trend is to eliminate the trend by differencing and data correction and to find an appropriate stationary model for the differenced series. Trend modeling by this approach requires: i) identifying the type of differencing required, ii) correcting the data for missing values, and/or, periodicities, iii) adjusting the data for seasonality and man-made intervention, and iv) identifying the appropriate orders of a stationary models for the differenced and corrected data.

A study was conducted at the Nevada Agricultural Experimental Station at the University of Nevada at Reno to explore the possibility of estimating the long term trend of a river water quality time series through the powerful tools provided by the SAS®/ETS software [7] for time series analysis such as PROC EXPAND, PROC ARIMA, and PROC MODEL. Water quality at most of the natural systems display a trend with time was the hypothesis lead to this study. SAS®/ETS software was used to successfully identify: a) autocorrelation functions (ACF), partial ACF, and inverse ACF across time, b) type of cross correlation relations among time series, c) type of differencing required to remove trend and seasonality, d) identifying model orders and fitness, e) diagnostic checking of the identified model and, for, f) residual analysis. All the graphics reported here are produced using the new experimental SAS® ODS GRAPHICS option in Version 9.13.

## OBJECTIVES

1. To examine the feasibility of using time series analysis to detect long-term water quality trend.
2. To explore the usefulness of the SAS®/ETS software for long-term water quality trend modeling.

## DATA INTEGRITY

### ORIGINAL TIME SERIES DATA

Total nitrate nitrogen concentration (tn) for the period from January 1980 to December 1998 was converted to SAS® data set from EXCEL. A PLOT of the original series is shown in Figure 1. Trend of the original series appear to be slightly increasing. However, this needs to be tested and confirmed through descriptive analysis and trend modeling.

### MISSING VALUES, AND PERIODICITIES

It is important to verify the periodicity of the time series before applying a trend model to it. For the same the observations should be dated correctly, and the data set sorted by date. Stochastic time series models require evenly spaced data without any missing values. The EXPAND procedure in SAS® corrects the data for missing values, and periodicities. TO=MONTH with PROC EXPAND can be used to interpolate the full set of time series converting the data frequency to monthly values and correcting for periodicities. The BEGINNING statement with the OBSERVED= indicates that the data are beginning-of-period values while the AVERAGE statement indicates that the data values represent period averages. A graphical display of the data corrected for missing values and periodicities requested by specifying the ODS GRAPHICS statement with the PLOT= option with PROC EXPAND in SAS® (SAS CODE 1) is shown in Figure 2.

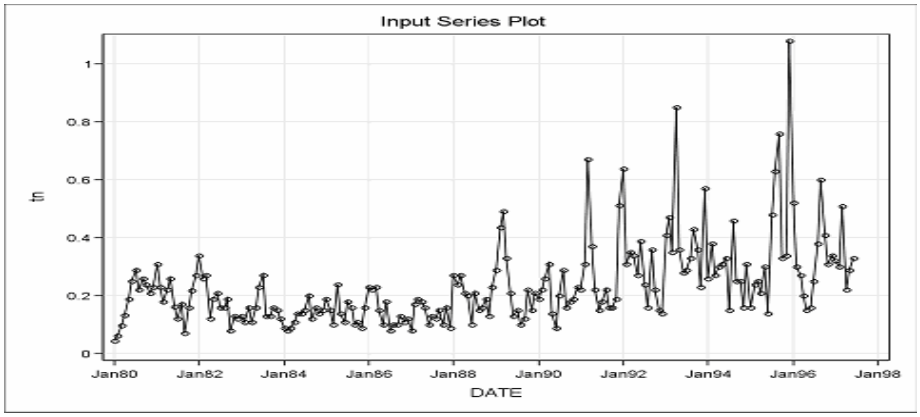


**Figure1: Original total nitrogen concentration**

SAS® CODE 1

```
PROC EXPAND DATA=TREND OUT=MONTHLY TO=MONTH PLOT=ALL;
     ID DATE ;
     CONVERT NTO=TN / OBSERVED=(BEGINNING, AVERAGE);
   RUN;
```

### SEASONALITY AND MAN MADE INTERVENTION MODELING

Most water quality time series exhibit variation that is annual in period (i.e., high in summer but lower in winter). Seasonality can be readily estimated or can be removed from the data to give de-seasonalized data if seasonality is not of interest. Summer and winter seasonality and any man-made intervention ($X_1$) in the data were included in the analysis as explanatory variables estimated as

$X_1$ = 'man made intervention'
summer = 'summer months'
winter = 'winter months'
$X_1$ = year >=1990
summer = ( 5 < mm < 11 ) * ( year > 1965 );
winter = ( year > 1965 ) - summer;

A useful class of models for time series is formed by combining MA and AR processes. An integrated mixed AR and MA process (ARIMA) contains p, AR terms, and q, MA terms and d orders of differencing and said to be an

ARIMA process of order (p+d, q). An ARIMA model was fitted to identify the influence of the summer, winter and, man-made intervention variables on tn through the PROC ARIMA procedure in SAS® and the significance of the parameters estimated were checked at the 5% level of significance (SAS® CODE 2).

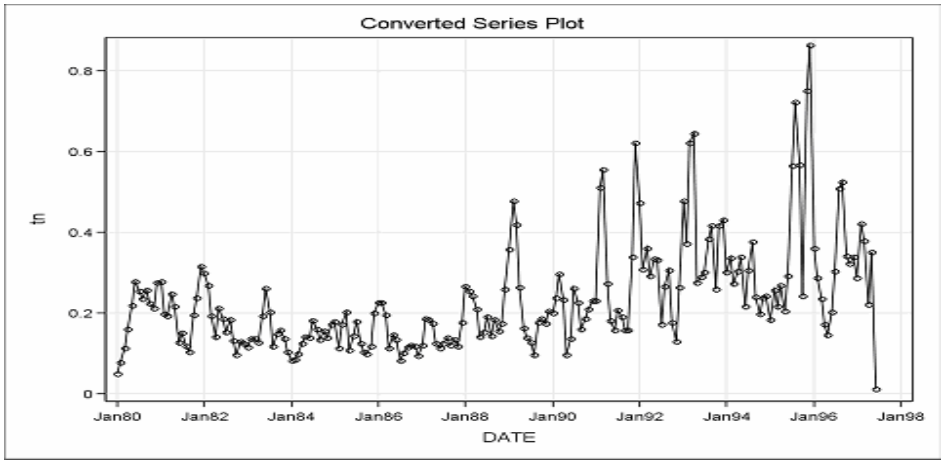


**Figure 2: Total nitrate nitrogen corrected for missing values**

AUTOCORRELATION AND OUTLIERS

Most water quality time series exhibit autocorrelation over time. In an autocorrelated time series positive deviations from a mean are followed by positive deviations and negative deviations by negative deviations or in the case of negative autocorrelation positive deviations are followed by negative and negative by positive. The SAS® procedure ARIMA produces autocorrelation function (ACF), partial ACF, and inverse ACF plots from which the presence of autocorrelation in the series can be identified (SAS CODE 2). Significant spikes in ACF, PACF, and IACF plots at specified frequencies indicate the presence of autocorrelation at the frequencies (Figure 3, Figure 4 and Figure 5). The SCAN option with the IDENTIFY statement in PROC ARIMA identifies the autoregressive and moving average model orders tentatively output as SCAN tables of (p+d, q) terms at 5% level of significance that can fit the original data. Vertices of right triangular regions in which all elements are insignificant at the 5% level of significance are selected as tentative autoregressive (AR) and moving average (MA) model orders (Table 1).

SAS® CODE 2

```
PROC ARIMA DATA=monthly;
      IDENTIFY VAR=tn(1,12) STATIONARITY=(ADF=(1,2,4,6,12))
      CROSSCORR=(  summer winter X1  ) SCAN;
      ESTIMATE P=(1 2 4 6 12)  Q=(1  12) NOINT
      INPUT=( summer  winter x1) PLOT METHOD=ML;
   RUN;
```

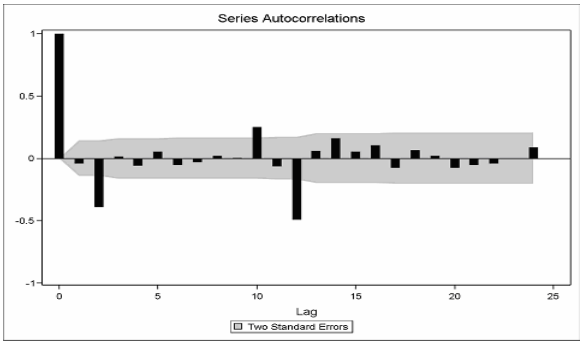


**Figure 3: ACF plot for model order identification**

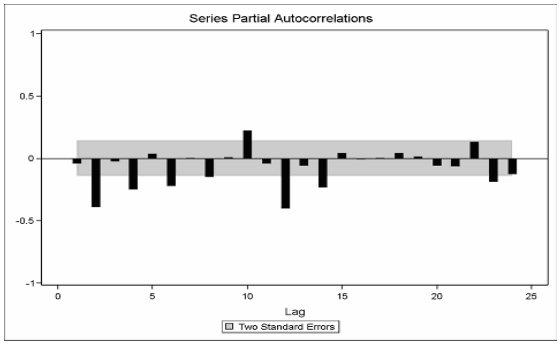


**Figure 4: PACF plot for model order identification**

PROC ARIMA also produces results of model diagnostic tests (Table 2) and maximum likelihood estimates of the specified model parameters with probabilities as part of the SAS® output (Table 7).

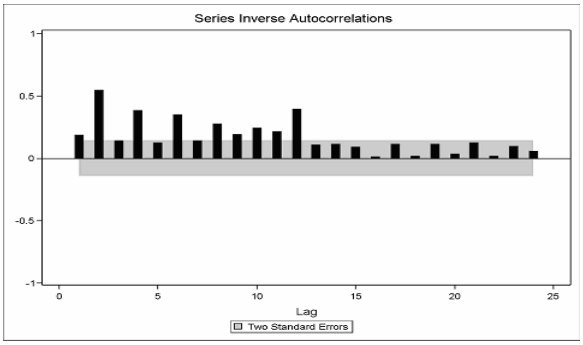


**Figure 5: IACF plots for model order identification**

The outlier is indeed an extreme observation, although, the converse is not necessarily true. On the balance of probabilities an observation beyond 3 standard deviation need to be highlighted for follow-up investigation to identify causes such as man-made intervention, and recording error. Only the most extreme observations (4 or more standard deviations from the mean) need to be excluded from the analysis. An intervention analysis was performed on the extreme values (3 standard deviations above mean).

**NON-STATIONARITY**

A time series ($X_t$, t (time)=0, ±1, ±2,….) is said to be stationary if it has statistical properties similar to those of the time-shifted series ($X_{t+h}$ , t=0, ±1, ±2,….) for each integer h. In practice most time series are non-stationary and has a trend and often the non-stationary component (trend) is of most interest. However, much of the probability theory of time series is concerned with stationary time series, and most of the stochastic time series models require that the data be stationary [3]. SAS® provides the formal test Dicky and Fuller to test for unit root non-stationarity with the ARIMA procedure (SAS® CODE 2). Since significant *p*+*d* terms (1, 2, 4, 6, and 12) were identified (Table 1), a unit root test is required to determine whether the autoregressive terms are unit root. Since moving average terms were present, with large autoregressive terms the Augmented Dickey-Fuller test for a unit root is appropriate. The SAS output of the Dicky and Fuller test results indicates that there is unit root non-stationarity in the data (Table 3).

**Table 1: SCAN table with Tentative AR and MA model orders**

| SCAN Probability Values | | | | | | |
|---|---|---|---|---|---|---|
| **Lags** | **MA 0** | **MA 1** | **MA 2** | **MA 3** | **MA 4** | **MA 5** |
| **AR 0** | <.0001 | <.0001 | 0.0001 | 0.0014 | 0.0053 | 0.0041 |
| **AR 1** | <.0001 | 0.5974 | 0.4622 | 0.9198 | 0.4292 | 0.8768 |
| **AR 2** | <.0001 | 0.9022 | 0.6106 | 0.9182 | 0.5683 | 0.8592 |
| **AR 3** | <.0001 | 0.0002 | 0.2400 | 0.8713 | 0.8930 | 0.7372 |
| **AR 4** | <.0001 | 0.8380 | 0.3180 | 0.7737 | 0.8913 | 0.6213 |
| **AR 5** | <.0001 | <.0001 | 0.0765 | 0.9142 | 0.1038 | 0.8103 |

**Table 2: Model diagnostic characteristics**

| Parameter | Values |
|---|---|
| **Variance Estimate** | 0.007875 |
| **Std Error Estimate** | 0.088742 |
| **AIC** | -374.673 |
| **SBC** | -351.691 |
| **Number of Residuals** | 197 |

**DIFFERENCING**

If a time series is non-stationary, it must be 'derived' to induce stationarity. By differencing a time series we 'derive' the function by which it is generated which removes the trend and any seasonal and non-seasonal cycles and periodicities found in the original series. Since significant autocorrelations were found at lag1 and at lag12 from ACF, PACF and IACF plots 1st and 12th differencing were performed on the data using the differencing capabilities provided with SAS®. PROC ARIMA differences the data when specified with the VAR= option in the IDENTIFY statement as: **IDENTIFY VAR=TN(1 12)** and performs the non-stationarity tests and estimation of model orders and parameters with probabilities on the differenced series. However, fitting seasonally adjusted ARIMA model requires the data be differenced using the DIF function prior to fitting the model as shown in SAS® CODE 3.

SAS® CODE 3

```
DATA monthly; SET monthly;
   tn1=DIF(tn);
   tn12=DIF12(tn1);
 RUN;
```

## TREND MODELING

Advanced ARIMA stochastic model [2] was fitted to the trend removed, seasonally adjusted water quality time series using the model orders identified through PROC ARIMA. An ARIMA model in general consists of a system of one or more linear equations that predict a value in a response time series as a linear combination of its own past values, past errors, and current and past values of other time series. The AR and MA macros in SAS® identifies and writes the equations of the system in SAS® programming language and solves the equations through the FIT option with PROC MODEL (SAS CODE 4) for the specified AR and MA model orders. However, the orders of the model must be identified as well the type of differencing required prior to fitting the model with PROC MODEL. PROC MODEL allows multiple input and output series and interaction among the input and output series provided as a linear equation through the MODEL = option and estimates the model parameters with their probabilities by the method specified with AR and MA macros. Original water quality variable can as well be tested for normality, and white noise using the NORMAL and WHITE options with the FIT statement in PROC MODEL (SAS® CODE 4). The normality (Table 4) and the white noise (Table 5) test results indicate that the original series is not-normal and not-white noise. Data were 1st and 12th differenced (SAS® CODE 3) prior to fitting the model to convert the series into stationary. The fitted model is shown in Table 6. The test statistics for the estimated individual parameters of the model were different from zero at 0.05, level of significance (Table 7). The parameter estimates produced from PROC MODEL was compared to the estimates from PROC ARIMA (Table 7). Estimates from PROC MODEL were significant and close in value for all the parameters identified by PROC ARIMA at the 5% level.

**SAS CODE 4**

```
PROC MODEL DATA=monthly; /* MODEL=MONTHMOD */
  tn12=0;
  %AR(tn12,12,,1 2 4 6 12,    M=ML)
  %MA(tn12,12,,1 12,  M=ML)  /* %MA always after %AR */
  FIT tn12 /PRL=LR  OUT=OUT1 OUTEST=SOL1 MAXITER=30 OUTALL CONVERGE =
            0.1 NORMAL WHITE;
ID DATE;

   RUN;
QUIT;
```

A plot of the PROC MODEL predicted data (solid lines) (Figure 6) along with original data (circles), and the 95 percent confidence band is shown in Figure 6. As predicted by the model there is a slightly increasing trend in the water quality (total nitrogen (tn)) series.

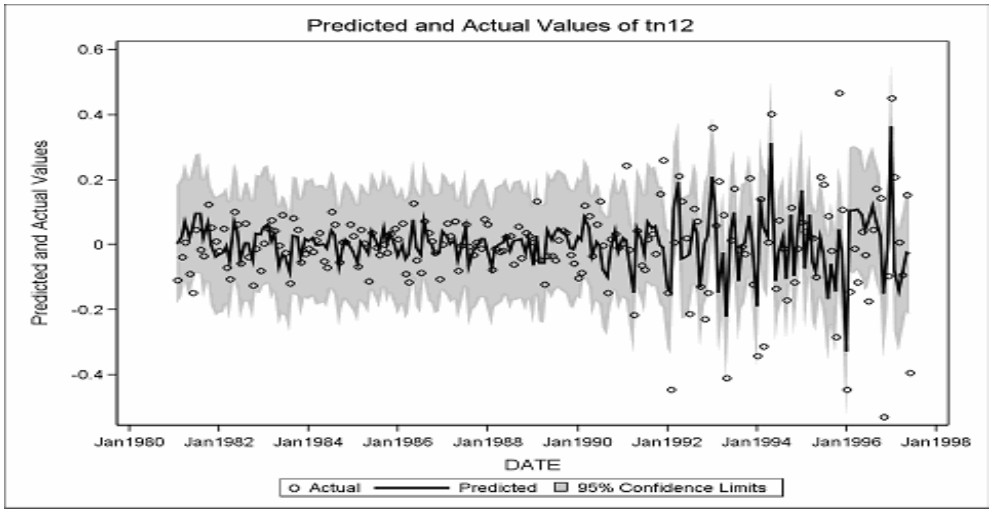


**Figure 6: PROC Model predicted plot of tn with original and 95% confidence band**

**Table 3: Augmented Dicky - Fuller unit root test results from PROC ARIMA**

| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
|---|---|---|---|---|---|---|---|
| Zero Mean | 1 | -487.826 | 0.0001 | -15.22 | <.0001 | | |
| | 2 | -526.531 | 0.0001 | -10.24 | <.0001 | | |
| | 4 | 1406.955 | 0.9999 | -8.26 | <.0001 | | |
| | 6 | 245.2934 | 0.9999 | -7.22 | <.0001 | | |
| | 12 | 51.6370 | 0.9999 | -6.54 | <.0001 | | |
| Single Mean | 1 | -487.920 | 0.0001 | -15.18 | <.0001 | 115.25 | 0.0010 |
| | 2 | -526.600 | 0.0001 | -10.22 | <.0001 | 52.22 | 0.0010 |
| | 4 | 1406.674 | 0.9999 | -8.24 | <.0001 | 33.94 | 0.0010 |
| | 6 | 245.1328 | 0.9999 | -7.20 | <.0001 | 25.92 | 0.0010 |
| | 12 | 51.5210 | 0.9999 | -6.52 | <.0001 | 21.29 | 0.0010 |
| Trend | 1 | -487.923 | 0.0001 | -15.14 | <.0001 | 114.64 | 0.0010 |
| | 2 | -526.374 | 0.0001 | -10.18 | <.0001 | 51.93 | 0.0010 |
| | 4 | 1429.085 | 0.9999 | -8.20 | <.0001 | 33.79 | 0.0010 |
| | 6 | 244.1137 | 0.9999 | -7.19 | <.0001 | 25.90 | 0.0010 |
| | 12 | 51.3196 | 0.9999 | -6.51 | <.0001 | 21.24 | 0.0010 |

**Table 4: Normality test results**

| Equation | Test Statistic | Value | Probability |
|---|---|---|---|
| tn12 | Shapiro-Wilk W | 0.88 | <.0001 |
| System | Mardia Skewness | 103.8 | <.0001 |
| | Mardia Kurtosis | 18.52 | <.0001 |
| | Henze-Zirkler T | 4.67 | <.0001 |

**Table 5: Heteroscdasticity test results**

| Equation | Test | Statistic | DF | Pr > ChiSq | Variables |
|---|---|---|---|---|---|
| tn12 | White's Test | 93.39 | 35 | <.0001 | Cross of all vars |

**Table 6: Summary of fitted model from PROC MODEL**

| Model Variables | tn12 |
|---|---|
| Parameters(Value) | tn12_l1(0) tn12_l2(0) tn12_l4(0) tn12_l6(0) tn12_l12(0) tn12_m1(0) tn12_m12(0) |
| Equations<br>tn12 = | F(tn12_l1, tn12_l2, tn12_l4, tn12_l6, tn12_l12, tn12_m1, tn12_m12) |

**Table 7: Model parameter estimates with probabilities**

**FROM PROC MODEL**

| Parameter | Estimate | Approx Std Err | t Value | Approx Pr > \|t\| | Label |
|---|---|---|---|---|---|
| tn12_l1 | -0.20769 | 0.0847 | -2.45 | 0.0152 | AR(tn12) tn12 lag1 parameter |
| tn12_l2 | -0.54463 | 0.0763 | -7.13 | <.0001 | AR(tn12) tn12 lag2 parameter |
| tn12_l4 | -0.3588 | 0.0831 | -4.32 | <.0001 | AR(tn12) tn12 lag4 parameter |
| tn12_l6 | -0.24797 | 0.0764 | -3.24 | 0.0014 | AR(tn12) tn12 lag6 parameter |
| tn12_l12 | -0.17174 | 0.0930 | -1.85 | 0.0663 | AR(tn12) tn12 lag12 parameter |
| tn12_m1 | -0.26337 | 0.0853 | -3.09 | 0.0023 | MA(tn12) tn12 lag1 parameter |
| tn12_m12 | 0.631945 | 0.0971 | 6.51 | <.0001 | MA(tn12) tn12 lag12 parameter |

**FROM PROC ARIMA**

| Parameter | Estimate | Standard Error | t Value | Approx Pr > \|t\| | Lag | Variable | Shift |
|---|---|---|---|---|---|---|---|
| MA1,1 | -0.38594 | 0.07505 | -5.14 | <.0001 | 1 | tn | 0 |
| MA1,2 | 0.61102 | 0.07328 | 8.34 | <.0001 | 12 | tn | 0 |
| AR1,1 | -0.24978 | 0.07968 | -3.13 | 0.0020 | 1 | tn | 0 |
| AR1,2 | -0.52358 | 0.07585 | -6.90 | <.0001 | 2 | tn | 0 |
| AR1,3 | -0.31532 | 0.08159 | -3.86 | 0.0002 | 4 | tn | 0 |
| AR1,4 | -0.22789 | 0.07441 | -3.06 | 0.0025 | 6 | tn | 0 |
| AR1,5 | -0.19554 | 0.08764 | -2.23 | 0.0268 | 12 | tn | 0 |

**RESIDUAL ANALYSIS**

PROC MODEL allows a complete analysis for residual assumptions such as autocorrelation, white noise and normality (SAS CODE 4). According to plots of residual ACF (Figure 7), partial ACF (Figure 8), and, inverse ACF (Figure 9) residuals are white noise and not-autocorrelated. According to the Q-Q plot of residuals (Figure 10) and the histogram of residuals (Figure 11) residuals of the model are normal. According residual Cook's D plot there are influential observations in the residuals (Figure 12). The OLS residual error summary is shown in Table 8.

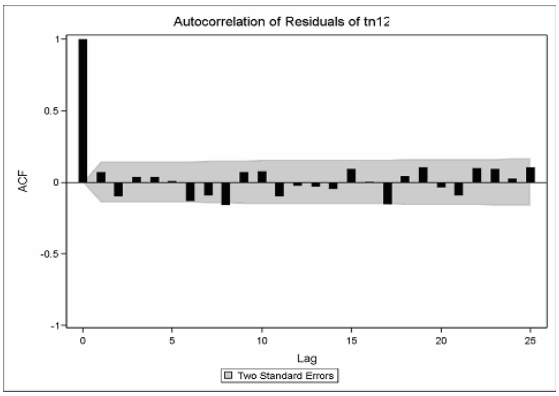


**Figure 7: Residual ACF plot**

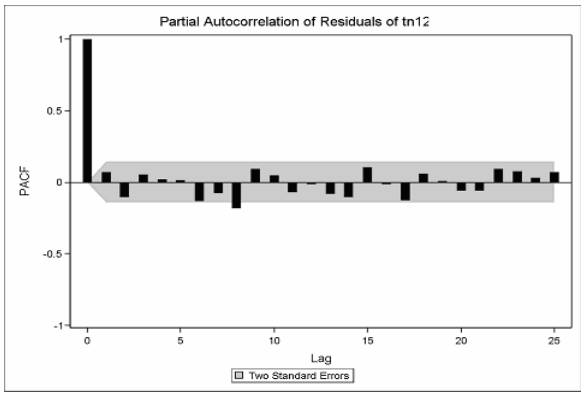


**Figure 8: Residual partial ACF plot**

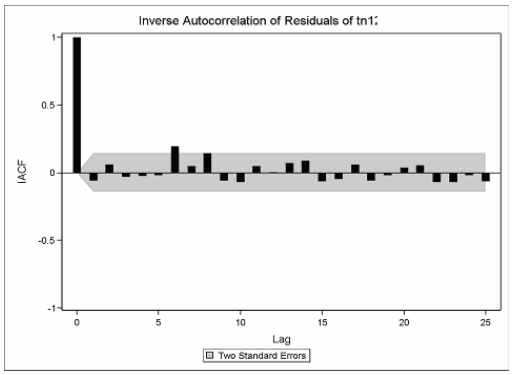


Figure 9: Residual inverse ACF

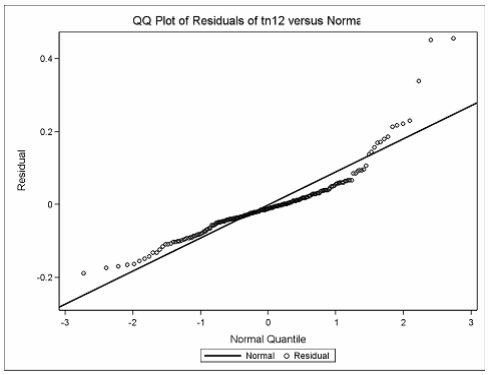


Figure 10: Q-Q plot of residuals

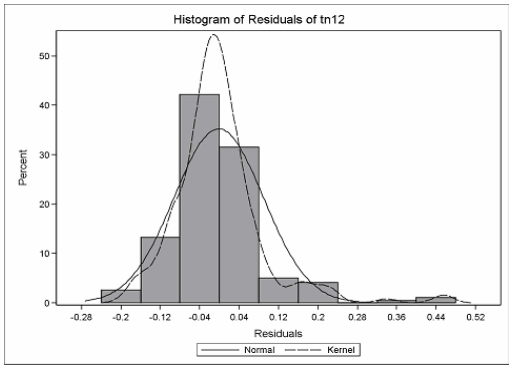


Figure 11: Histogram of residuals

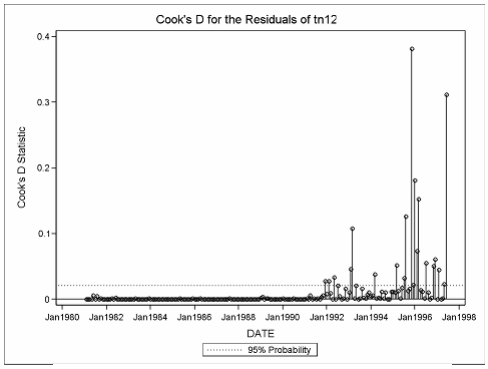


Figure 12: Residual Cook's D plot

**Table 8: Nonlinear OLS Summary of Residual Errors**

| Equation | DF Model | DF Error | SSE | MSE | Root MSE | R-Square | Adj R-Sq |
|---|---|---|---|---|---|---|---|
| tn12 | 7 | 190 | 2.2891 | 0.0120 | 0.1098 | 0.3471 | 0.3265 |
| RESID.tn12 | | 190 | 1.6086 | 0.00847 | 0.0920 | | |

## THE UCM PROCEDURE

Another useful approach to modeling trend is to decompose the original series into its components, such as trend, seasonality, and irregular residuals and analyze the components separately. This is useful when the different individual components are of interest. The UCM procedure in SAS® (SAS® CODE 5) decomposes the series into seasonal, trend and irregular error allowing for further analysis of the individual

SAS® CODE 5

```
PROC UCM data=monthly;
      ID date INTERVAL=month;
      MODEL tn;
      IRREGULAR;
      LEVEL;
      SLOPE VARIANCE=0 NOEST;
      SEASON LENGTH=12 TYPE=TRIG PLOT=SMOOTH;
      ESTIMATE BACK=12 PLOT=(RESIDUAL NORMAL ACF);
      FORECAST BACK=12 LEAD=12 PLOT=(FORECAST DECOMP);
   RUN;
```

components. The model also forecasts decomposed series for trend after removing seasonality and the irregular variations. ACF, PACF, and, IACF plots of the residuals can be obtained from the ESTIMATE statement of the model (SAS® CODE 5). Plots of the PROC UCM produced components: seasonal, trend and error of the series are shown in Figure 13, Figure 14, and Figure 15. The UCM forecasted series show slightly increasing trend (Figure 16). The error component of the model is normal (Figure 17) and with no autocorrelation (Figure 18). The overall model is significant at the 5% level (Table 9).

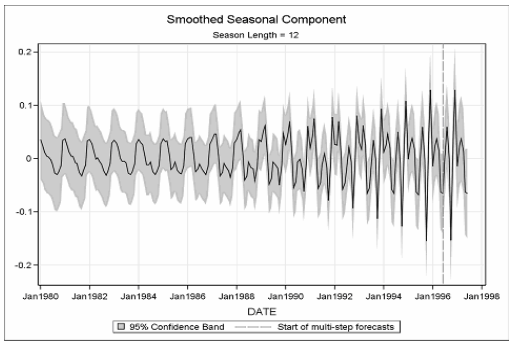


**Figure 13: Smoothed seasonal component**

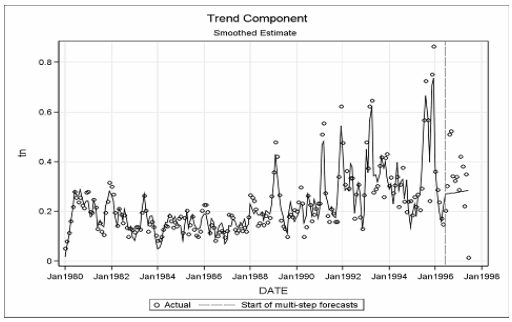


**Figure 14: Smoothed trend component**

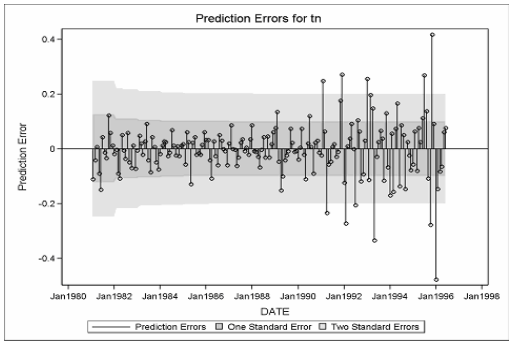


**Figure 15: Error of the series**

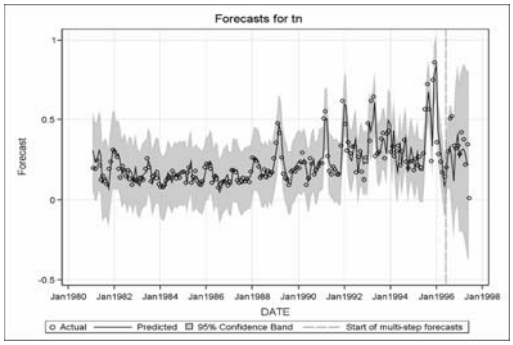


**Figure 16: The forecasted series**

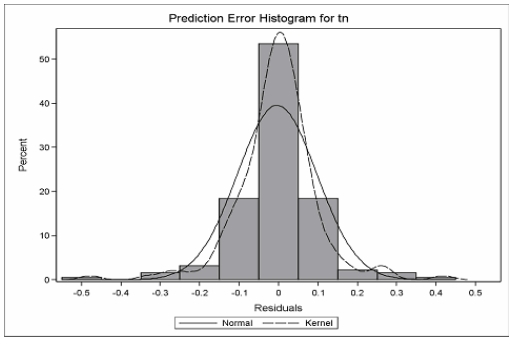


**Figure 17: Residual error histogram**

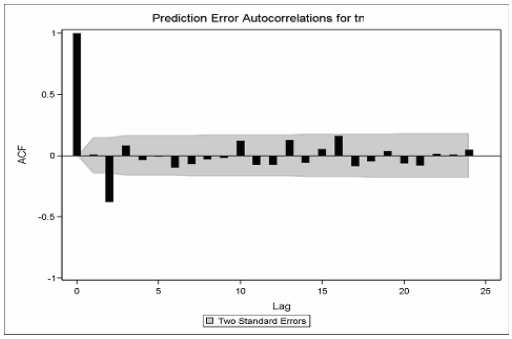


**Figure 18: Residual autocorrelogram**

**Tabl e 9: Parameter estimates of the final model from PROC UCM**

| Component | Parameter | Estimate | Approx Std Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| **Level** | Error Variance | 0.00727 | 0.0009815 | 7.41 | <.0001 |
| **Season** | Error Variance | 0.00000541 | 3.1575E-6 | 1.71 | 0.0869 |

## CONCLUSIONS

Seasonnaly differenced ARIMA modeing using SAS® is a successful approach to model the trend of the water quality in rivers. Original time series data were, non-stationary, seasonal and had strong positive autocorrelation. Data also contained missing values. Data were first and $12^{th}$ differenced, corrected for missing values and outliers. Summer and winter variation as well as man-made intervention were included in the analysis as indicator variables and tested for their significance. Appropriate ARIMA model orders were identified through the ARIMA procedure in SAS®. An ARIMA model was fitted to the series using the identified model orders using the MODEL procedure in SAS®. Series normality and white noise were tested as well as a complete residual analysis was performed using the MODEL procedure in SAS®. Parameter estimates, of the ARIMA model was significant at 5% level. Predicted water quality showed slightly increasing trend. Model predicted residuals were normal, white noise and showed no autocorrelation. SAS®/ETS provides very powerful features to model time series analysis of water quality trend.

## REFERENCES


[1] Akaike, H., 1974, *A new look at the statistical model identification*, IEEE trans. Autom. Control, AC-19:716-723

[2] Box, G.E.P., and G.M. Jenkins, 1976, *Time series Analysis Forecasting and Control*, $2^{nd}$ ed., Holden-Day, San Francisco, California.

[3] Brockwell, P.J., R.A. Davis, 1996, *Introduction to time series and forecasting*, Springer, New York.

[4] Fuller, F.C., Jr., and C.P. Tsokos, 1971, *Time Series analysis of water pollution data*. Virginia Polytechnic Institute and State University, Blackybury, Virginia.

[5] Hipel, K.W., 1985, *Time series analysis in* perspective, Water Resources Bull., 21: 609-624.

[6] Salas, J.D., and J.T.B. Obeysekera, 1988, ARIMA models Identification of Hydrologic Time Series. Water Resources Research. Vol. 18, No. 4: 1011-1021.

[7] SAS®/ETS Software Applications Guide 1, *Time Series Modeling and Forecasting, Financial Reporting and loan Analysis*, Version 6, First Edition, SAS Institute Inc., Cary, NC 27513.

[8] SAS®/ETS Users Guide, 1993, Version 6, Second Edition, SAS Institute Inc., Cary, NC 27513.


## CONTACT INFORMATION

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


| | | | |
|---|---|---|---|
| Name: | Anpalaki J. Ragavan | Name: | Dr. George C. Fernandez, |
| Enterprise: | University of Nevada | Enterprise: | University of Nevada |
| Address: | 3952, Clear Acre Lane, # 276, Reno, NV 89512, USA. | Address: | CABNR/204 Reno, NV 89557 , U.S.A. |
| Work phone: | (775)-784-4433 | Work phone: | (775)-784-4206 |
| Fax: | (775)-784-1080 | Fax: | (775) 784-1342 |
| Email: | ragavan@unr.edu | Email: | gcjf@unr.edu |
| Web: | None | Web: | http://www.ag.unr.edu/gf |