Application 1: Analysis of an Uncensored Constituent using a Predefined Model

Dave Lorenz

July 20, 2015

This example illustrates the format of the input datasets and the format of the calls to build a predefined rating curve model and to make estimates in rloadest. A predefined model that describes the linear relation between the log of constituent load and log streamflow is built and used for load estimation.

The data used in this application were collected from the Illinois River at Marseilles, Illinois (p. 257, Helsel and Hirsch, 2002). Ninety six water-quality samples for total phosphorus collected between November 1974 and April 1985 are used in conjunction with observed streamflow data to build the regression model and the calibration data are used to estimate loads as was dome in Helsel and Hirsch (2002) and Runkel and others (2004).

Part 2 illustrates the diagnostic graphs that can be used to improve the model and offers a step-by-step approach to building a calibrated model.

```
> # Load the rloadest package and the data
> library(rloadest)
> data(app1.calib)
> head(app1.calib)
```

```
       DATES TIMES  FLOW Phosphorus
1 1974-11-13  1200  6860       1.40
2 1974-12-16  1200  6640       0.69
3 1975-01-14  1200  9870       0.74
4 1975-02-12  1200  8690       1.00
5 1975-03-17  1200 11300       0.55
6 1975-04-15  1200  9010       0.50
```

1 Build the Model

The `loadReg` function is used to build the rating-curve model for constituent load estimation. The basic form of the call to `loadReg` is similar to the call to `lm` in that it requires a formula and data source. The response variable in the formula is the constituent concentration, which is converted to load per day (flux) based on the units of concentration and the units of flow. The `conc.units`, `flow.units`, and `load.units` arguments to `loadReg` define the conversion. For these data, the concentration units (`conc.units`) are "mg/L", the flow units are "cfs" (the default), and the load units for the model are "kg" (also the default). If `conc.units` is not set, they are assumed to be "mg/L" and a warning is issued. Two additional pieces of information are required for `loadReg`—the names of the flow column and the dates column. A final option, the station identifier, can also be specified.

Predefined models can easily be constructed using the `model` function as the response variable. For the call to `loadReg`, only the model number is needed—the `loadReg` automatically constructs the required input. This example uses model number 1. The model numbers match the terms in Runkel and others (2004), but the order is different—decimal time terms precede seasonal time terms.

```
> # Create the load model.
> app1.lr <- loadReg(Phosphorus ~ model(1), data = app1.calib, flow = "FLOW",
+                    dates = "DATES", conc.units="mg/L",
+                    station="Illinois River at Marseilles, Ill.")
```

2 Print the Model Report

An abbreviated form of the model report can be printed simply by typing the name of the model object (`app1.lr` in this case) in the R Console window. A more complete form that closely matches the output from LOADEST can be obtained by using the `print` function as shown below.

```
> print(app1.lr, brief=FALSE, load.only=FALSE)

                      LOADEST
       A Program to Estimate Constituent Loads
U.S. Geological Survey, Version for R 0.1 (June, 2013)
-----------------------------------------------------

Station: Illinois River at Marseilles, Ill.
Constituent: Phosphorus

-------------------------------------------------------------------------
    Constituent Output File Part Ia: Calibration (Load Regression)
-------------------------------------------------------------------------

           Number of Observations: 96
Number of Uncensored Observations: 96
           Center of Decimal Time: 1979.601
                  Center of ln(Q): 9.1779
                Period of record: 1974-11-13 to 1985-04-15

Model Evaluation Criteria Based on AMLE Results
-----------------------------------------------

  model   AIC  SPCC
1     1 68.96 76.65
Model # 1 selected

Selected Load Model:
--------------------

Phosphorus ~ model(1)

where:
   Phosphorus is the constituent load in log(kg/d)
and model 1 has these variables:
   lnQ is ln(Q) - center of ln(Q)

Model coefficients:
            Estimate Std. Error z-score p-value
```

```
(Intercept)    9.336    0.03750   248.9       0
lnQ            0.761    0.05399    14.1       0

AMLE Regression Statistics
Residual variance: 0.1152
R-squared: 67.88 percent
G-squared: 109 on 1 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9893
  p-value = 0.1095
Serial Correlation of Residuals: 0.2303


Comparison of Observed and Estimated Loads
------------------------------------------
  The summary statistics and bias diagnostics presented below are based
on a comparison of observed and estimated loads for all dates/times within
the calibration data set. Although this comparison does not directly
address errors in load estimation for unsampled dates/times, large
discrepancies between observed and estimated loads are indicative of a
poor model fit. Additional details and warnings are provided below.

Note: The comparison that follows uses a concentration equal to 1/2 the
detection limit when an observation is censored. The summary stats and
bias diagnostics are therefore slightly inaccurate for censored datasets.

      Summary Stats: Loads in kg/d
-------------------------------------------
     Min  25%  50%   75%   90%   95%   Max
Est 4510 6990 9320 13000 18500 25100 49000
Obs 2430 5960 8390 12100 21500 33800 77400

Bias Diagnostics
----------------
 Bp: -1.977 percent
PLR: 0.9802
  E: 0.6105

where:
   Bp Load Bias in Percent
      Positive (negative) values indicate over (under) estimation.
      ***The model should not be used when the + or - bias exceeds 25%***
  PLR Partial Load Ratio
      Sum of estimated loads divided by sum of observed loads.
      Values greater than 1 indicate over estimation.
```

```
      Values less than 1 indicate under estimation.
    E Nash Sutcliffe Efficiency Index
      E ranges from -infinity to 1.0
      E = 1; a perfect fit to observed data.
      E = 0; model estimates are as accurate as the mean of observed data.
      E < 0; the observed mean is a better estimate than the model estimates.

NOTE: Additional information on model calibration is included in the
      residual diagnostic plots. users should conduct a thorough
      residuals analysis. Example residual plots are shown in figures
      7, 8, 9, and 17 of the LOADEST documentation (Runkel et al., 2004).

-------------------------------------------------------------------------
    Constituent Output File Part Ia: Calibration (Concentration Regression)
-------------------------------------------------------------------------

Model # 1 selected

Selected Concentration Model:
-----------------------------

Phosphorus ~ model(1)

where:
   Phosphorus is the constituent concentration in log(mg/L)
and model 1 has these variables:
   lnQ is ln(Q) - center of ln(Q)

Model coefficients:
            Estimate Std. Error z-score p-value
(Intercept)  -0.7366    0.03750 -19.643       0
lnQ          -0.2390    0.05399  -4.427       0

AMLE Regression Statistics
Residual variance: 0.1152
R-squared: 17.25 percent
G-squared: 18.18 on 1 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9893
  p-value = 0.1095
Serial Correlation of Residuals: 0.2303


Comparison of Observed and Estimated Concentrations
---------------------------------------------------
```

```
  The summary statistics and bias diagnostics presented below are based
on a comparison of observed and estimated concentrations for all dates/times
within the calibration data set. Although this comparison does not directly
address errors in concentration estimation for unsampled dates/times, large
discrepancies between observed and estimated concentrations are indicative of
a poor model fit. Additional details and warnings are provided below.

Note: The comparison that follows uses a concentration equal to 1/2 the
detection limit when an observation is censored. The summary stats and
bias diagnostics are therefore slightly inaccurate for censored datasets.

      Summary Stats: Concentrations in mg/L
-----------------------------------------------
     Min  25%  50%  75%  90%  95%  Max
Est 0.32 0.49 0.55 0.60 0.65 0.67 0.69
Obs 0.23 0.39 0.50 0.64 0.82 0.95 1.40

Bias Diagnostics
----------------
 Bp: -0.2367 percent
PCR: 0.9976
  E: 0.1491

where:
   Bp Concentration Bias in Percent
      Positive (negative) values indicate over (under) estimation.
      ***The model should not be used when the + or - bias exceeds 25%***
  PCR Partial Concentration Ratio
      Sum of estimated concentrations divided by sum of observed concentrations.
      Values greater than 1 indicate over estimation.
      Values less than 1 indicate under estimation.
    E Nash Sutcliffe Efficiency Index
      E ranges from -infinity to 1.0
      E = 1; a perfect fit to observed data.
      E = 0; model estimates are as accurate as the mean of observed data.
      E < 0; the observed mean is a better estimate than the model estimates.

NOTE: Additional information on model calibration is included in the
      residual diagnostic plots. users should conduct a thorough
      residuals analysis. Example residual plots are shown in figures
      7, 8, 9, and 17 of the LOADEST documentation (Runkel et al., 2004).
```

Aside, from the cosmetic differences, there will be some differences in the actual numeric output. Major differences are listed below.

LOADEST prints a modified form of AIC and SPCC, whereas the AIC and SPCC computed by this version are consistent with AIC and BIC

computed for the same model using different methods, like simple linear regression (using `lm`) in this case of no censoring.

The format for the model output matches the general format for regression model output in `R` rather than the format in LOADEST. It is expected that users of rloadest will be familiar with the general format for regression model output in `R`.

This version prints G-squared, which is the test statistic for the overall model fit, and it's attained p-value.

Finally, the summary statistics of loads and concentrations are based on the units defined in the call to `loadReg` rather than the specified output in LOADEST.

3 Estimate Loads

Unlike LOADEST, rloadest requires to the user to build the rating-curve model before estimating loads. For this application, we will follow the general layout of LOADEST and estimate loads directly from the model created earlier in this application.

The `predLoad` function is used to estimate loads. It estimates loads in units per day, which is referred to as flux in rloadest. The arguments for `predLoad` are `fit`, the model output from `loadReg`; `newdata`, the estimation dataset; `load.units`, the load units for the estimates, which are taken from the model output if not specified; `by`, a character string indicating how to aggregate the load estimates; `seopt`, how to compute the standard error of the load; `allow.incomplete`, a logical value that indicates whether or not to allow incomplete periods to be estimated; and `print`, indicating whether to print a summary.

Unlike the `predict` function in base `R`, `newdata` is required. The columns in `newdata` must match the column names in the calibration dataset. For predefined models, the column names for dates and flow must match.

The `by` argument must be "unit," "day,", "month," "water year," "calendar year," "total," or the name of a grouping column in `newdata`. The "unit" option is not available in version 0.1.

The argument `allow.incomplete` is not fully implemented in version 0.1.

Application 1 in LOADEST uses the identical data for estimation as was used for calibration. This application will use the same dataset. The call in the `R` code below simply prints the results and discards the data frame that is calculated.

```
> predLoad(app1.lr, newdata = app1.calib, load.units="tons", by="total",
+          print=TRUE)

------------------------------------------------------------------------
Constituent Output File Part IIa: Estimation (test for extrapolation)
 Load Estimates for 1974-11-13 to 1985-04-15
------------------------------------------------------------------------

Streamflow Summary Statistics
-----------------------------------------

The maximum estimation data set streamflow does not exceed the maximum
calibration data set streamflow. No extrapolation is required.


------------------------------------------------------------
Constituent Output File Part IIb: Estimation (Load Estimates)
```

```
 Load Estimates for 1974-11-13 to 1985-04-15
------------------------------------------------------------------

Flux Estimates, in tons/d, using AMLE
---------------------------------------

  Period Ndays     Flux   Std.Err       SEP      L95      U95
1  total    96 12.37659 0.5113722 0.7320215 11.00375 13.87217
```

4 Part 2, Diagnostic Plots

The rloadest package contains a `plot` function that creates diagnostic plots of the load model. Most often the user will just enter `plot(app1.lr)` (for this example) in the R Console window to generate the full suite of plots, but this example application will generate each plot individually. And, in general, the user will not need to set up a graphics device. But for this vignette, the graphics device must be set up for each graph.

Figure 1 is related to figure 7 in Runkel and others (2004) because there is only a single explanatory variable. Figure 1 shows the AMLE regression line as a dashed line and the solid line is a LOWESS smooth curve. The LOWESS curve agrees very well with the regression line.

```
> # setSweave is required for the vignette.
> setSweave("app1_01", 5, 5)
> plot(app1.lr, which=1, set.up=FALSE)
> graphics.off()
```

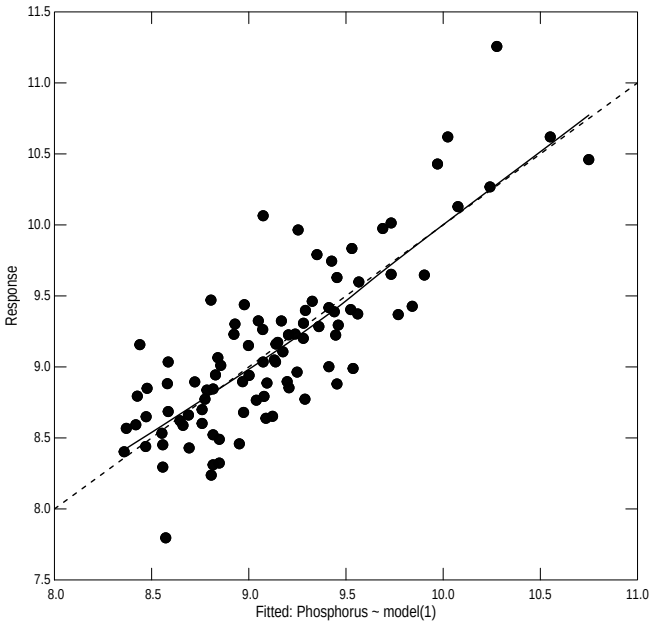

Figure 1. The rating-curve regression model.

Figure 2 is the same as figure 8 in Runkel and others (2004).The horizontal dashed line is at zero and the solid line is the LOWESS smooth. The LOWESS smooth is very close to the zero line and indicates no lack of fit.

```
> # setSweave is required for the vignette.
> setSweave("app1_02", 5, 5)
> plot(app1.lr, which=2, set.up=FALSE)
> graphics.off()
```

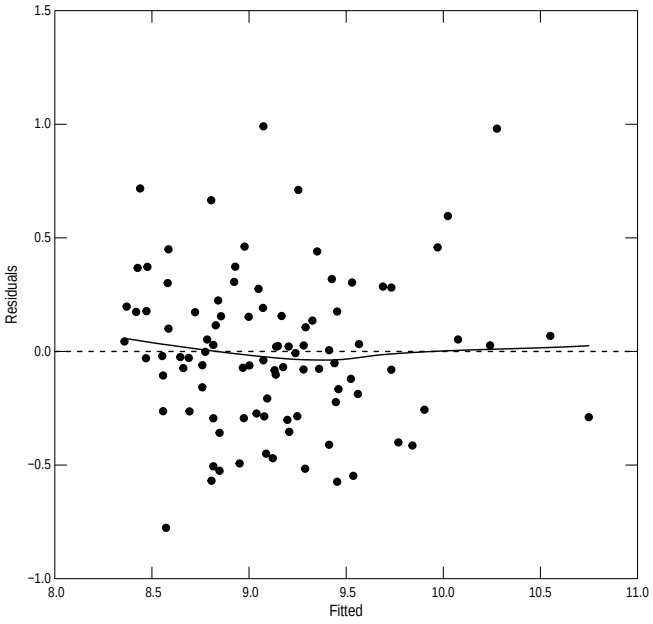

Figure 2. The residuals versus fit for the regression model.

Figure 3 is a scale-location (S-L) graph that is a useful graph for assessing heteroscedasticity of the residuals. The horizontal dashed line is the expected value of the square root of the absolute value of the residuals and the solid line is the LOWESS smooth. Even though there is a small slope in the LOWESS line, it is not enough to cause concern for unequal variance in the estimates.

```
> # setSweave is required for the vignette.
> setSweave("app1_03", 5, 5)
> plot(app1.lr, which=3, set.up=FALSE)
> graphics.off()
```

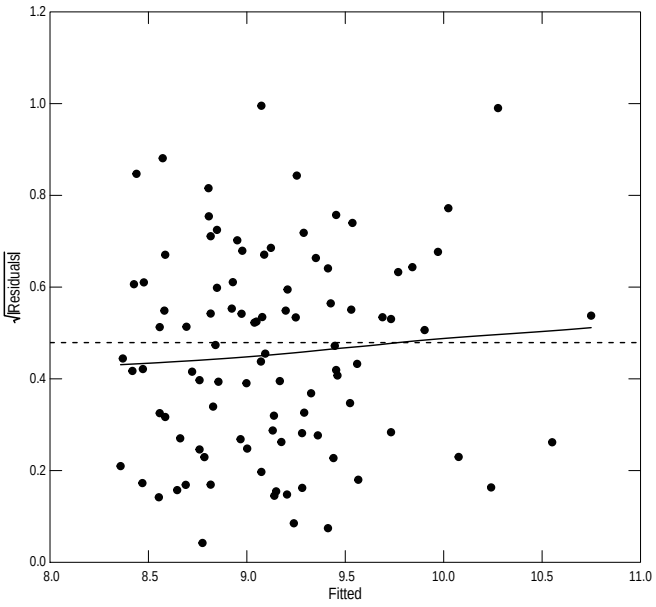

Figure 3. The scale-location graph for the regression model.

The correlogram in figure 4 is a adaptation of the correlogram from time-series analysis, which deals with regular samples. The horizontal dashed line is the zero value and the solid line is a kernel smooth rather than a LOWESS line. The kernel smooth gives a better fit in this case. The solid line should be very close to the horizontal line. In this case, because the solid line is consistently above the horizontal line for more than 1 year, we have concern for a lack of fit over time—a linear time term should be added to the model. There is also a slight regular pattern with a higher line at 0 and 1 and a low line at about 0.5. This might suggest a seasonal lack of fit.

```
> # setSweave is required for the vignette.
> setSweave("app1_04", 5, 5)
> plot(app1.lr, which=4, set.up=FALSE)
> graphics.off()
```

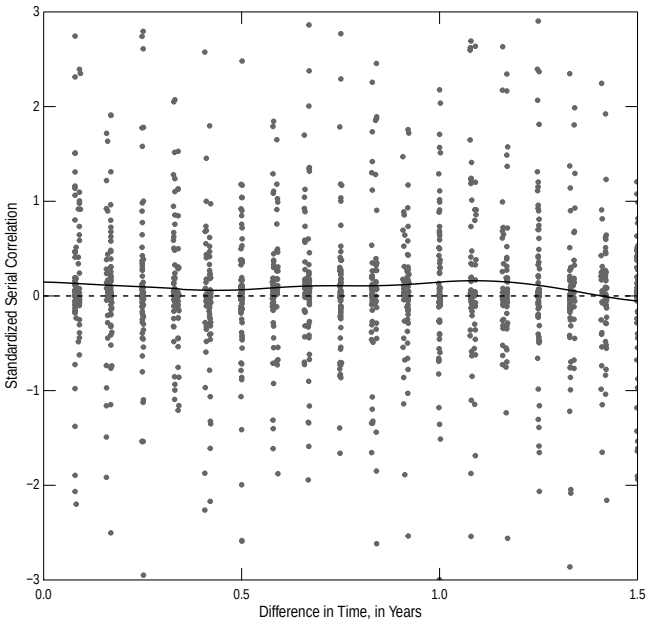

Figure 4. The correlogram from the regression model.

Figure 5 is the same as figure 9 in Runkel and others (2004), except that figure 5 shows the standardized residuals, which are assumed to have a standard deviation of 1. The solid line is the theoretical fit of mean of 0 and standard deviation of 1. The visual appearance of figure 5 confirms the results of the PPCC test in the printed output—the residuals are reasonably normal in distribution.

```
> # setSweave is required for the vignette.
> setSweave("app1_05", 5, 5)
> plot(app1.lr, which=5, set.up=FALSE)
> graphics.off()
```

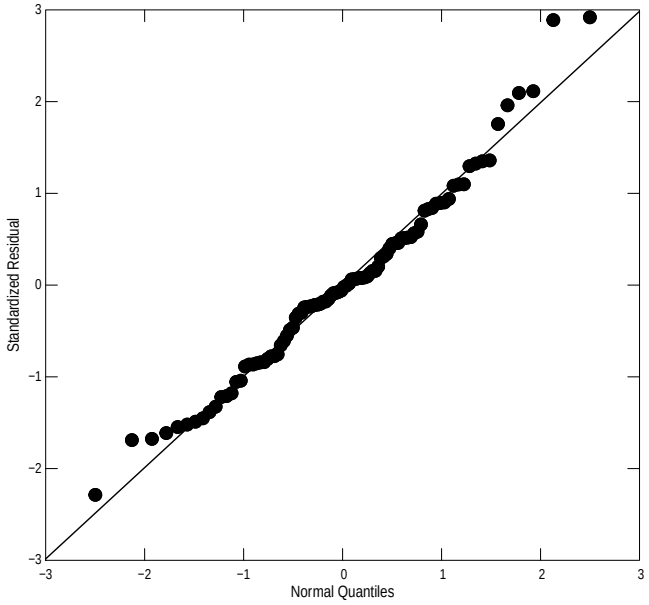

Figure 5. The Q-normal plot of the residuals.

Figure 6 is an extended box plot—a truncated box plot, at the 5 and 95 percentiles that shows the individual values larger than the 95th percentile and smaller than the 5th percentile. The box plots in figure 6 show the distributions of the actual and estimated loads. The appearance of these box plots agrees with what is expected—the range of the estimated values is a little smaller than the range of the actual,because of "regression to the mean," and the location is similar, the median and quartiles match reasonably well. This figure confirms the bias diagnostics section of the printed report.

```
> # setSweave is required for the vignette.
> setSweave("app1_06", 5, 5)
> plot(app1.lr, which=6, set.up=FALSE)
> graphics.off()
```

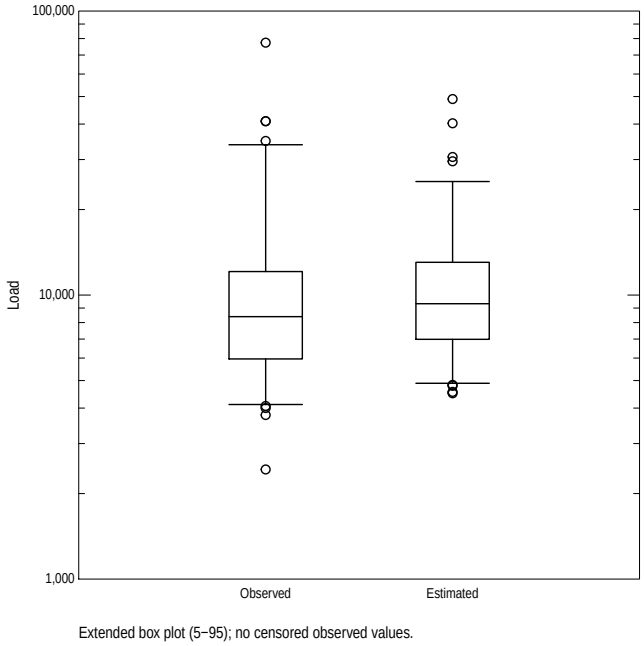

Figure 6. Box plot comparing estimated and observed values.

5 Add Linear and Seasonal Time to the Model

The correlogram (fig. 4) suggested that at least one reason for the relatively large value for the serial correlation (0.2303) being relatively large is a lack of fit over time and possibly a seasonal component. Model number 7 includes linear flow, linear time and seasonal time. Note the brevity of the brief version (default) of the report. The correlogram, fig. 7, does not show a regular pattern that might indicate any lack of fit over time. The remaining diagnostic plots are not displayed in this vignette.

```
> # Create and print the revised load model.
> app1.lr7 <- loadReg(Phosphorus ~ model(7), data = app1.calib, flow = "FLOW",
+                     dates = "DATES", conc.units="mg/L",
+                     station="Illinois River at Marseilles, Ill.")
> print(app1.lr7)

*** Load Estimation ***

Station: Illinois River at Marseilles, Ill.
Constituent: Phosphorus

           Number of Observations: 96
Number of Uncensored Observations: 96
           Center of Decimal Time: 1979.601
                  Center of ln(Q): 9.1779
                Period of record: 1974-11-13 to 1985-04-15

Selected Load Model:
--------------------

Phosphorus ~ model(7)

Model coefficients:
            Estimate Std. Error z-score p-value
(Intercept)  9.33402    0.03277 284.827  0.0000
lnQ          0.84910    0.05310  15.991  0.0000
DECTIME     -0.05998    0.01077  -5.566  0.0000
sin.DECTIME -0.02418    0.04677  -0.517  0.5957
cos.DECTIME  0.15091    0.04197   3.596  0.0004

AMLE Regression Statistics
Residual variance: 0.0789
R-squared: 78.71 percent
G-squared: 148.5 on 4 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
```

```
  r = 0.9823
  p-value = 0.0132
Serial Correlation of Residuals: -0.0064

Variance Inflation Factors:
              VIF
lnQ         1.413
DECTIME     1.033
sin.DECTIME 1.332
cos.DECTIME 1.063

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
------------------------------------------------
     Min  25%  50%   75%   90%   95%   Max
Est 4370 6340 8580 13500 18300 27400 59000
Obs 2430 5960 8390 12100 21500 33800 77400

Bias Diagnostics
----------------
 Bp: -1.143 percent
PLR: 0.9886
  E: 0.7002

> # setSweave is required for the vignette.
> setSweave("app1_07", 5, 5)
> plot(app1.lr7, which=4, set.up=FALSE)
> graphics.off()
```

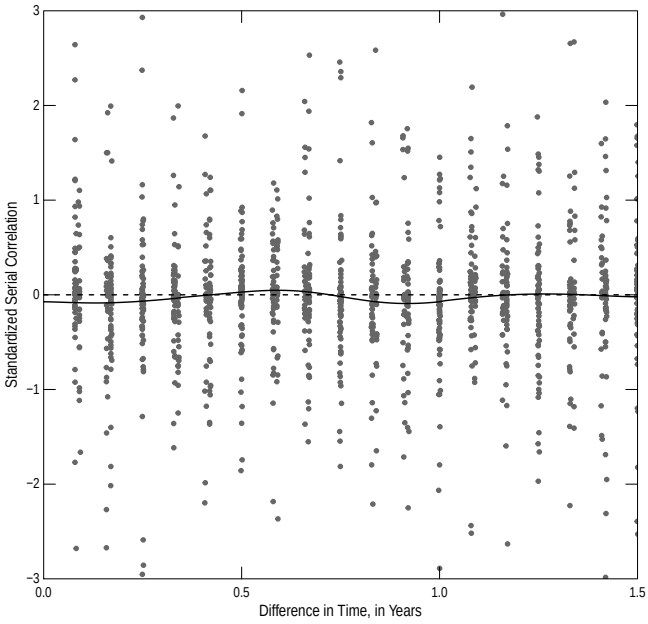

Figure 7. The correlogram from the revised regression model.