Incorporating Hysteresis in a Load Model

Dave Lorenz

July 26, 2017

This example illustrates how to set up and incorporate hysteresis into a load model. Hysteresis occurs when concentrations (and thereby loads) are different on rising and falling limbs of an event hydrograph for the same magnitude of streamflow.

This example extends the analysis of Garrett (2012) of nitrate plus nitrite loads in the Boyer River at Logan, Iowa, USGS gaging station 06609500. The original time frame for the analysis was for water years 2004 through 2008. This example extends the analysis through water year 2012. Loads will not be estimated from this model—it demonstrates only the building and evaluation phases.

```
> # Load the necessary packages and the data
> library(rloadest)
> library(dataRetrieval)
> # Get the QW data
> Boyer <- "06609500"
> BoyerQW <- importNWISqw(Boyer, "00631",
+   begin.date="2003-10-01", end.date="2012-09-30")
> # Now the Daily values data
> BoyerQ <- readNWISdv(Boyer, "00060", startDate="2003-10-01",
+   endDate="2012-09-30")
> BoyerQ <- renameNWISColumns(BoyerQ)
```

1 Compute the Hysteresis Variables and Merge the Data

There is no direct diagnostic test to determine if hysteresis is important. The best way to assess any hysteresis is to compute the hysteresis metric and plot residuals against that metric. The first section of code will replicate that effort, beginning with the "best" predefined model rather than simply replicating the model used by Garrett (2012). Garrett used a 1-day lag. This example will use that and add a 3-day lag metric. The function that computes the metric is `hysteresis` in **smwrBase**.

```
> # Compute the hysteresis metrics.
> BoyerQ <- transform(BoyerQ,
+   dQ1 = hysteresis(log(Flow), 1),
+   dQ3 = hysteresis(log(Flow), 3))
> # Rename the date column in QW so that the data can be merged
> names(BoyerQW)[2] <- "Date"
> # Merge the data
> BoyerData <- mergeQ(BoyerQW, FLOW=c("Flow", "dQ1", "dQ3"),
+   DATES="Date", Qdata=BoyerQ, Plot=F)
> # Create the initial model
> Boyer.lr <- selBestModel("NO2PlusNO3.N", BoyerData,
+   flow="Flow", dates="Date", station=Boyer)
> print(Boyer.lr)

*** Load Estimation ***

Station: 06609500
Constituent: NO2PlusNO3.N

           Number of Observations: 101
Number of Uncensored Observations: 101
           Center of Decimal Time: 2008.411
                  Center of ln(Q): 6.4937
                 Period of record: 2004-03-12 to 2012-09-10

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

  model    AIC   SPCC   AICc
1     1 103.57 111.42 103.82
2     2  64.40  74.86  64.82
3     3 103.76 114.22 104.17
4     4 103.93 117.00 104.56
5     5  60.70  73.77  61.33
```

```
6     6  64.93  80.62  65.83
7     7 104.34 120.03 105.23
8     8  61.57  79.88  62.78
9     9  57.24  78.16  58.81
Model # 9 selected

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

NO2PlusNO3.N ~ model(9)

Model coefficients:
             Estimate Std. Error   z-score p-value
(Intercept)  9.522594   0.050343 189.15421  0.0000
lnQ          0.916603   0.031300  29.28439  0.0000
lnQ2        -0.092397   0.013286  -6.95429  0.0000
DECTIME     -0.026111   0.012144  -2.15018  0.0277
DECTIME2    -0.015283   0.006199  -2.46541  0.0119
sin.DECTIME  0.104600   0.048919   2.13824  0.0285
cos.DECTIME  0.002443   0.046596   0.05243  0.9567

AMLE Regression Statistics
Residual variance: 0.09464
R-squared: 94.34 percent
G-squared: 290.1 on 6 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9646
  p-value = 1e-04
Serial Correlation of Residuals: 0.1058

Variance Inflation Factors:
              VIF
lnQ         1.605
lnQ2        1.046
DECTIME     1.030
DECTIME2    1.307
sin.DECTIME 1.231
cos.DECTIME 1.167

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
---------------------------------------------
    Min  25%  50%   75%   90%   95%    Max
Est 221 3040 8470 16100 35100 51200 121000
```

```
Obs 166 3010 8230 16600 29400 42800 178000

Bias Diagnostics
----------------
 Bp: -2.817 percent
PLR: 0.9718
  E: 0.7912
```

The model selected was number 9, which includes second-order flow and time terms and the first-order seasonal terms. Reviewing the table of model evaluation criteria, model number 2 had a very small value for AIC and the second smallest for SPPC. Model number 2 would have been the equivalent starting model in Garrett (2012), including only the second-order flow terms. The printed results indicate good bias statistics, but the PPCC p-value is much less than 0.05. The next section of code illustrates plotting the residuals and hysteresis metrics. The 1-day lag appears to fit better than the 3-day lag.

```
> # residuals and hysteresis
> setSweave("graph01", 6, 8)
> AA.lo <- setLayout(num.rows=2)
> setGraph(1, AA.lo)
> AA.pl <- xyPlot(BoyerData$dQ1, residuals(Boyer.lr),
+   ytitle="Residuals", xtitle="1-day lag hysteresis",
+   xaxis.range=c(-1,3))
> addSLR(AA.pl)
> setGraph(2, AA.lo)
> AA.pl <- xyPlot(BoyerData$dQ3, residuals(Boyer.lr),
+   ytitle="Residuals", xtitle="3-day lag hysteresis",
+   xaxis.range=c(-1,3))
> addSLR(AA.pl)
> dev.off()

null device
          1
```

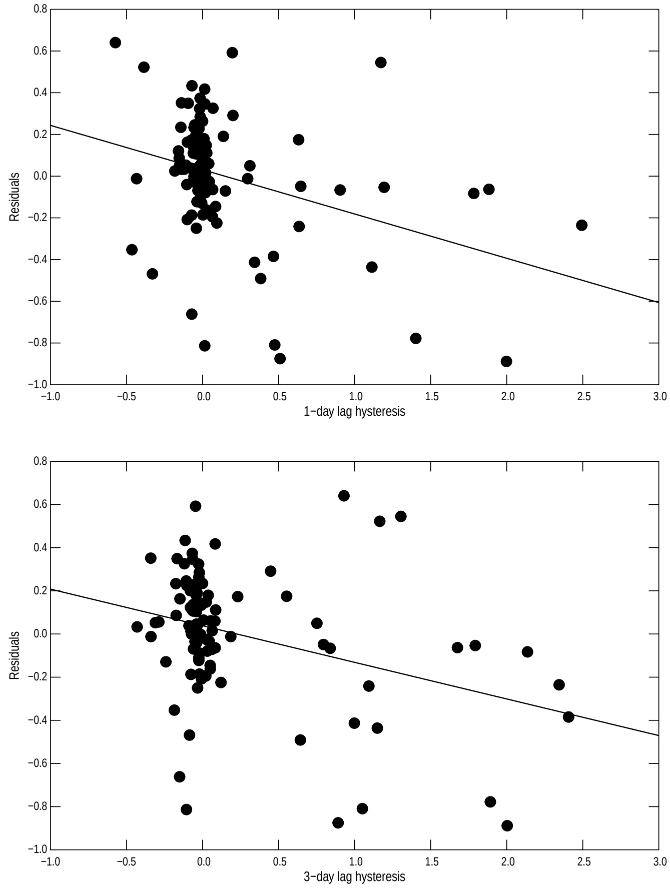

Figure 1. The residuals versus hysteresis metrics.

2 Construct and Validate the Hysteresis Model

The model...

```
> # Construct the model
> Boyer.lr <- loadReg(NO2PlusNO3.N ~ quadratic(log(Flow)) +
+   quadratic(dectime(Date)) + fourier(Date) + dQ1, data=BoyerData,
+   flow="Flow", dates="Date", station=Boyer)
> print(Boyer.lr)

*** Load Estimation ***

Station: 06609500
Constituent: NO2PlusNO3.N

           Number of Observations: 101
Number of Uncensored Observations: 101
           Center of Decimal Time: 2008.411
                  Center of ln(Q): 6.4937
                 Period of record: 2004-03-12 to 2012-09-10

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

NO2PlusNO3.N ~ quadratic(log(Flow)) + quadratic(dectime(Date)) +
    fourier(Date) + dQ1

Model coefficients:
                                    Estimate Std. Error  z-score
(Intercept)                         9.546076   0.045687 208.9442
quadratic(log(Flow))(6.49366)1      0.994584   0.032678  30.4362
quadratic(log(Flow))(6.49366)2     -0.079044   0.012313  -6.4196
quadratic(dectime(Date))(2008.41)1 -0.037313   0.011211  -3.3283
quadratic(dectime(Date))(2008.41)2 -0.009762   0.005713  -1.7088
fourier(Date)sin(k=1)               0.094337   0.044187   2.1350
fourier(Date)cos(k=1)               0.024549   0.042295   0.5804
dQ1                                -0.320703   0.067628  -4.7422
                                   p-value
(Intercept)                         0.0000
quadratic(log(Flow))(6.49366)1      0.0000
quadratic(log(Flow))(6.49366)2      0.0000
quadratic(dectime(Date))(2008.41)1  0.0007
quadratic(dectime(Date))(2008.41)2  0.0772
fourier(Date)sin(k=1)               0.0279
fourier(Date)cos(k=1)               0.5456
```

```
dQ1                                   0.0000

AMLE Regression Statistics
Residual variance: 0.07703
R-squared: 95.44 percent
G-squared: 312 on 7 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9652
  p-value = 1e-04
Serial Correlation of Residuals: 0.0863

Variance Inflation Factors:
                                   VIF
quadratic(log(Flow))(6.49366)1     2.150
quadratic(log(Flow))(6.49366)2     1.104
quadratic(dectime(Date))(2008.41)1 1.079
quadratic(dectime(Date))(2008.41)2 1.364
fourier(Date)sin(k=1)              1.234
fourier(Date)cos(k=1)              1.181
dQ1                                1.510

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
---------------------------------------------
    Min  25%  50%   75%   90%   95%    Max
Est 215 3000 8400 16400 30700 45700 181000
Obs 166 3010 8230 16600 29400 42800 178000

Bias Diagnostics
----------------
 Bp: 0.2131 percent
PLR: 1.002
  E: 0.9312
```

The printed output shows an improved model, but the PPCC test still indicates lack of normality. A review of the linearity of the explanatory variables indicates the need for second-order seasonal terms (figure 1). The sine term is also nonlinear, but not shown. The second order linear time term will be dropped because it is not significant at the 0.05 level.

```
> # Plot the overall fit, choose "fourier(Date)cos(k=1)"
> setSweave("graph02", 6, 6)
> plot(Boyer.lr, which="fourier(Date)cos(k=1)", set.up=FALSE)
> dev.off()
```

```
null device
          1
```

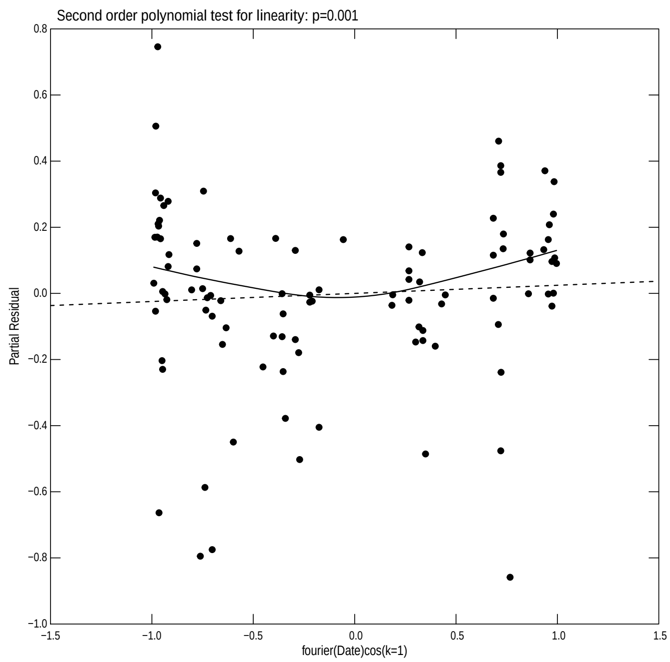

Figure 2. The diagnostic plot for the linearity of a seasonal term.

Construct the revised model that drops the second order time term and adds the second order seasonal terms. The diagnostic plots follow.

```
> # Construct the revised model
> Boyer.lr <- loadReg(NO2PlusNO3.N ~ quadratic(log(Flow)) +
+   dectime(Date) + fourier(Date, 2) + dQ1, data=BoyerData,
+   flow="Flow", dates="Date", station=Boyer)
> print(Boyer.lr)

*** Load Estimation ***

Station: 06609500
Constituent: NO2PlusNO3.N

           Number of Observations: 101
Number of Uncensored Observations: 101
```

```
          Center of Decimal Time: 2008.411
                 Center of ln(Q): 6.4937
                Period of record: 2004-03-12 to 2012-09-10

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

NO2PlusNO3.N ~ quadratic(log(Flow)) + dectime(Date) + fourier(Date,
    2) + dQ1

Model coefficients:
                               Estimate Std. Error z-score
(Intercept)                    87.79724   21.30386   4.121
quadratic(log(Flow))(6.49366)1  1.00848    0.02737  36.842
quadratic(log(Flow))(6.49366)2 -0.08300    0.01165  -7.123
dectime(Date)                  -0.03899    0.01060  -3.676
fourier(Date, 2)sin(k=1)        0.08242    0.04082   2.019
fourier(Date, 2)cos(k=1)        0.05428    0.03893   1.394
fourier(Date, 2)sin(k=2)       -0.06361    0.03709  -1.715
fourier(Date, 2)cos(k=2)        0.12670    0.03861   3.281
dQ1                            -0.34407    0.06304  -5.458
                               p-value
(Intercept)                     0.0000
quadratic(log(Flow))(6.49366)1  0.0000
quadratic(log(Flow))(6.49366)2  0.0000
dectime(Date)                   0.0002
fourier(Date, 2)sin(k=1)        0.0364
fourier(Date, 2)cos(k=1)        0.1461
fourier(Date, 2)sin(k=2)        0.0746
fourier(Date, 2)cos(k=2)        0.0008
dQ1                             0.0000

AMLE Regression Statistics
Residual variance: 0.06958
R-squared: 95.93 percent
G-squared: 323.3 on 8 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9656
  p-value = 2e-04
Serial Correlation of Residuals: 0.0449

Variance Inflation Factors:
                                 VIF
quadratic(log(Flow))(6.49366)1 1.670
quadratic(log(Flow))(6.49366)2 1.094
```

```
dectime(Date)                    1.069
fourier(Date, 2)sin(k=1)         1.165
fourier(Date, 2)cos(k=1)         1.107
fourier(Date, 2)sin(k=2)         1.038
fourier(Date, 2)cos(k=2)         1.033
dQ1                              1.453

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
---------------------------------------------
    Min  25%  50%   75%   90%   95%    Max
Est 195 3420 8460 15900 33900 46300 193000
Obs 166 3010 8230 16600 29400 42800 178000

Bias Diagnostics
----------------
 Bp: 0.2313 percent
PLR: 1.002
  E: 0.9379
```

A complete review of the partial residual graphs is not included in this example. Only the partial residual for `fourier(Date)cos(k=1)` is shown to show that the linearity has improved. No partial residual plot indicates a serious lack of linearity.

```
> # Plot the residual Q-normal graph.
> setSweave("graph03", 6, 6)
> plot(Boyer.lr, which = "fourier(Date, 2)cos(k=1)", set.up=FALSE)
> dev.off()

null device
          1
```

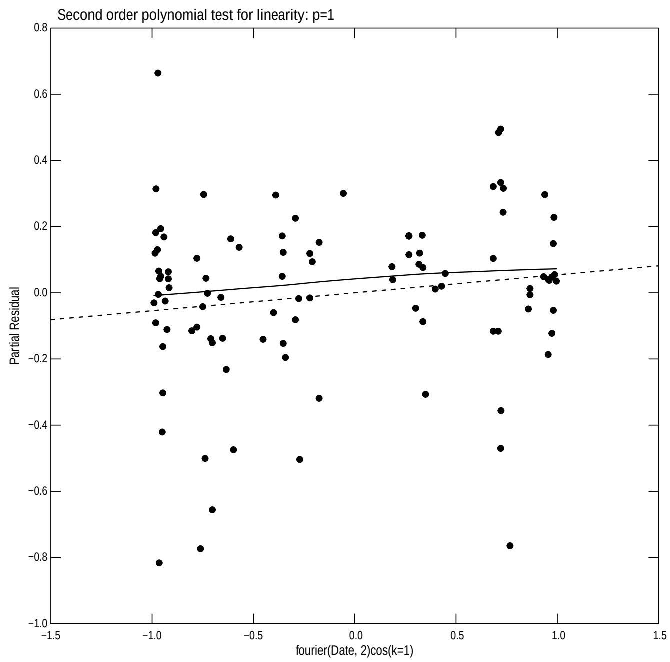

Figure 3. The partial residual for `fourier(Date, 2)cos(k=1)` graph.

The correlogram indicates a much better seasonal fit and no long-term lack of fit.

```
> # Plot the overall fit, choose plot number 2.
> setSweave("graph04", 6, 6)
> plot(Boyer.lr, which = 4, set.up=FALSE)
> dev.off()

null device
          1
```

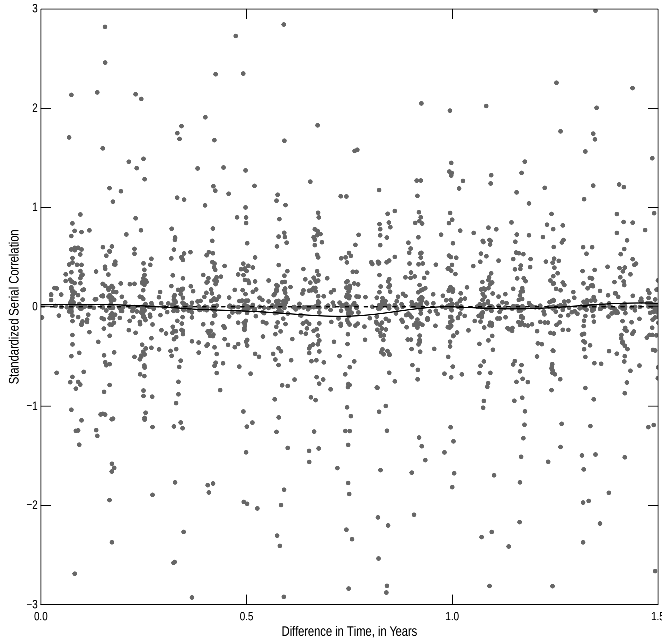

Figure 4. The correlogram for the revised model.

For this model, the S-L plot is shown. It shows a slight decrease in heteroscedasticity as the fitted values increase. The decrease is small and likely drive by the 3 largest fitted values and there is very little visual decrease in scatter as the fitted values increase.

```
> # Plot the S-L grpah.
> setSweave("graph05", 6, 6)
> plot(Boyer.lr, which = 3, set.up=FALSE)
> dev.off()

null device
          1
```

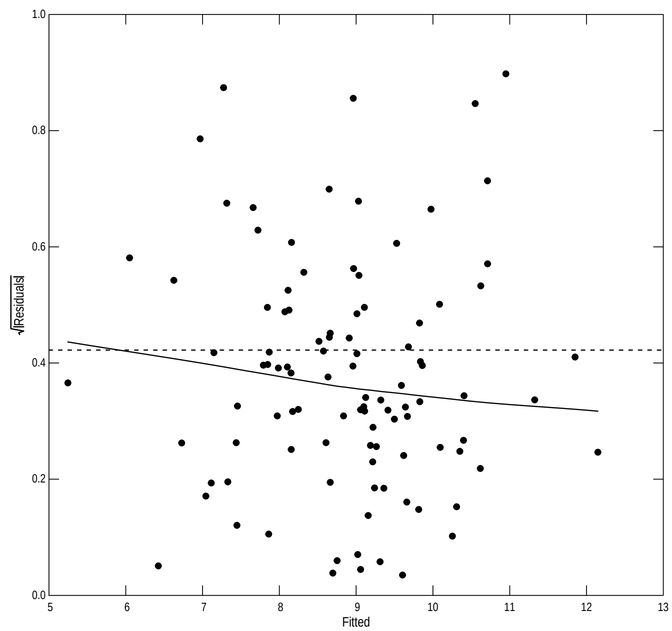

Figure 5. The S-L graph for the revised model.

The Q-normal graph shows a non-normal distribution, but without distinctive skewness or kurtosis.

```
> # Plot the residual Q-normal graph.
> setSweave("graph06", 6, 6)
> plot(Boyer.lr, which = 5, set.up=FALSE)
> dev.off()

null device
          1
```

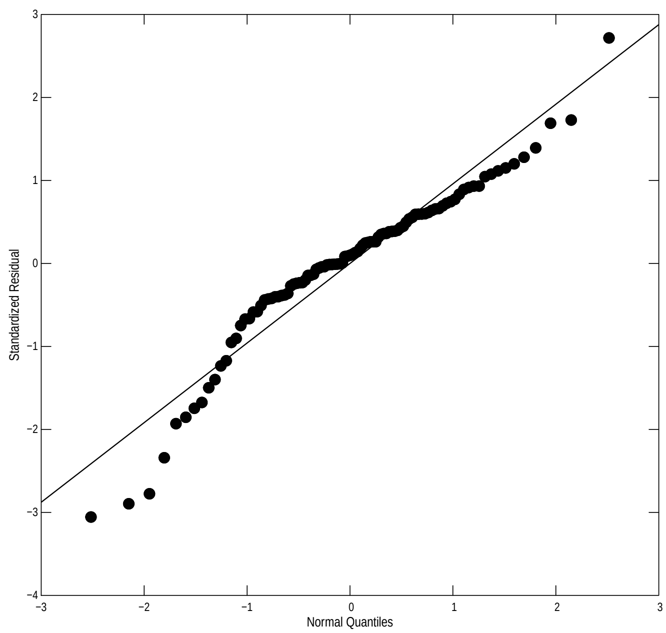

Figure 6. The residual Q-normal graph.

The non-normality of the residuals poses at least three potential problems: (1) lack of confidence in the confidence limits for regression parameter estimates, (2) bias in the back-transformation corrections for the mean load, and (3) lack of confidence in the confidence intervals of the load estimate. For the first potential problem, the confidence intervals of the parameter estimates and their significance are not critical to load estimation and the non-normality is not so large that this is a major concern (Greene, 2012). Several factors must be considered to address the second potential problem—the magnitude of the bias correction and the measured bias. The actual bias correction factor (BCF) used is AMLE, which cannot be directly computed, but can be estimated by using the MLE BCF. The MLE BCF for this model is `exp(0.06958/2) = 1.0354`. For non-normal data, Duan's smoothing BCF is often used (Helsel and Hirsch, 2002); the value for that BCF is `mean(exp(residuals(Boyer.lr))) = 1.0302`. The BCFs are very similar, which suggests that the bias from the back transform could be small. That is confirmed by the very small value for the percent bias in the printed report (0.2405). The third potential problem and no way to address it other than by using bootstrapping methods. Any reported confidence intervals on loads or fluxes should be treated as approximate.

References

[1] Garrett, J.D., 2012, Concentrations, loads, and yields of select constituents from major tributaries of the Mississippi and Missouri Rivers in Iowa, water years 2004???2008: U.S. Geological Survey Scientific Investigations Report 2012???-5240, 61 p.

[2] Greene, W.H., 2012, Econometric analysis, seventh edition: Upper Saddle River, New. Jersey, Prentice Hall, 1198 p.

[3] Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.