2 The Analytic Process

```
load('.RData')
```

2.1 Data Sets

Analyses of surface water chemistry uses three streams (called B, J, and S), pH, and date as explanatory (predictor) variables with Na, K, Mg, Ca, Cl, and SO4 as response variables. The R[1] code to perform these analyses is in Appendix 1.«»= ls() @

2.1.1 Stream B

The data available for stream B are in Appendix A.

The first step in analyzing water chemistry data for CWA compliance is reading it into the analytical software.

```
b <- read.table("./stream-b.dat", header = TRUE, sep = ",")
```

Next, check that the data are what you expect to see and convert the stream name from a factor to a character string and dates from factors to dates. The stream name is retained to identify the data from it when comparing among different streams.

```
b$stream <- as.character(b$stream)
b$sampdate <- as.Date(b$sampdate)
str(b)

## 'data.frame': 657 obs. of  4 variables:
##  $ stream  : chr  "B" "B" "B" "B" ...
##  $ sampdate: Date, format: "1992-03-27" "1992-03-27" ...
##  $ param   : Factor w/ 7 levels "Ca","Cl","K",..: 2 6 7 2 6 7 2 6 7 2 ...
##  $ quant   : num  4 33 8.43 4 32 8.46 4 31 8.43 6 ...
```

[1] R is an open-source language for statistics and data analyses available at no cost from the project's web site http://www.r-project.org/.

The above table shows the structure of the data in a R data.frame format. This format allows data of different types (here, character strings, dates, factors, and floating point numbers) to co-exist. We see that there are 657 observations in this data set. Another way of examining the data is to look at a summary of each of the data columns. We see that the earliest data date is 1992-03-07 and the latest date is 2011-08-23[2], and the number of observations for each chemical constituent.

```
summary(b)

##     stream             sampdate          param         quant
##  Length:657         Min.   :1992-03-27   Ca : 63   Min.   :  1.00
##  Class :character   1st Qu.:1994-08-18   Cl :142   1st Qu.:  6.52
##  Mode  :character   Median :1996-05-14   K  : 21   Median :  8.46
##                     Mean   :1999-02-12   Mg : 64   Mean   : 40.07
##                     3rd Qu.:2003-11-24   Na : 34   3rd Qu.: 50.00
##                     Max.   :2011-08-23   SO4:155   Max.   :784.00
##                                          pH :178
```

It is important to learn if there are missing values in this data set because some statistical models cannot produce accurate results if missing data (represented in R as "NA") are included. A script is used to check for these missing values.

```
show.col.with.na(b)

## Error in eval(expr, envir, enclos):  could not find function "show.col.with.na"
```

And it turns out there are no missing values in this data set.

Another look at the data set is provided by the summary:

[2]Using the International Standards Organization (ISO) format for dates ensures that they always sort in the correct order.