Data Development Procedure

### Sources and Structure of Raw Data

Two data sources were used as the basis for generating sample ecological data to be used in the expert elicitation process. Both datasets were generated by the North Carolina Department of Environmental Quality (DEQ). The first source is a data set that includes a wide range of ecological measurements across 68 creeks and rivers within the Haw River and Upper Neuse River watersheds, and was used as the source of measurements for five biophysical metric variables: fecal coliform, specific conductance, total nitrogen, total phosphorus, and turbidity.

## Rows: 6  
## Columns: 31  
## $ ...1 <dbl> 129, 189, 200, 22…  
## $ MonitoringLocationIdentifier <chr> "21NCCOALITIONS-J…  
## $ OrganizationIdentifier <chr> "21NCCOALITIONS",…  
## $ OrganizationFormalName <chr> "NC DWQ Monitorin…  
## $ ActivityIdentifier <chr> "21NCCOALITIONS-L…  
## $ ActivityTypeCode <chr> "Field Msr/Obs", …  
## $ Year <dbl> 2009, 2009, 2009,…  
## $ Month <dbl> 12, 10, 12, 7, 9,…  
## $ Day <dbl> 28, 8, 28, 2, 13,…  
## $ ExactDate <dbl> 2009.99, 2009.77,…  
## $ ActivityStartTime.Time <dttm> 1899-12-31 09:58…  
## $ ActivityDepthHeightMeasure.MeasureValue <dbl> 0.1, 0.1, 0.1, 0.…  
## $ ActivityDepthHeightMeasure.MeasureUnitCode <chr> "m", "m", "m", "m…  
## $ ProjectIdentifier <chr> "LNBA", "LNBA", "…  
## $ SampleCollectionMethod.MethodIdentifier <chr> "USEPA", "GRAB", …  
## $ SampleCollectionMethod.MethodIdentifierContext <chr> "USEPA", "21NCCoa…  
## $ SampleCollectionMethod.MethodName <chr> "USEPA", "GRAB", …  
## $ SampleCollectionEquipmentName <chr> "Unknown", "Water…  
## $ CharacteristicName <chr> "Temperature, wat…  
## $ ResultSampleFractionText <lgl> NA, NA, NA, NA, N…  
## $ ResultMeasureValue <dbl> 6.80, 92.00, 4.70…  
## $ ResultMeasure.MeasureUnitCode <chr> "deg C", "cfu/100…  
## $ ResultCommentText <lgl> NA, NA, NA, NA, N…  
## $ ResultAnalyticalMethod.MethodIdentifier <chr> NA, "9222D", "213…  
## $ ResultAnalyticalMethod.MethodIdentifierContext <chr> NA, "APHA", "APHA…  
## $ ResultAnalyticalMethod.MethodName <chr> NA, "Membrane fil…  
## $ MethodDescriptionText <chr> NA, "https://www.…  
## $ MonitoringLocationName <chr> "Falling Crk at S…  
## $ LatitudeMeasure <dbl> 35.32239, 35.3223…  
## $ LongitudeMeasure <dbl> -78.12815, -78.12…  
## $ watershed <chr> "UpperNeuse", "Up…

The second data source is a set of macroinvertibrate data collected from creeks and rivers in the Upper Neuse River watershed and the Cape Fear River watershed, of which the Haw River is a major tributary. This data set served as source of biotic index measurement data.

## Rows: 1,234  
## Columns: 43  
## $ `OBJECTID \*` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, …  
## $ `Shape \*` <chr> "Point", "Point", "Point", "Point",…  
## $ FID\_Macro <dbl> 1140, 1141, 1142, 2623, 1627, 1628,…  
## $ SiteID <chr> "BB223", "BB223", "BB500", "BB235",…  
## $ CC\_Num <dbl> 3201, 4351, 10432, 6279, 7685, 8182…  
## $ Waterbody <chr> "JAMES CR", "JAMES CR", "JAMES CR",…  
## $ Location <chr> "NR SR 2028", "NR SR 2028", "OFF SR…  
## $ Date\_ <dttm> 1984-02-23, 1986-03-05, 2008-06-30…  
## $ Sample\_Type <chr> "Full Scale", "Full Scale", "Full S…  
## $ Bioclass <chr> "Good", "Good", "Excellent", "Not R…  
## $ Latitude <dbl> 35.16111, 35.16111, 35.18722, 35.19…  
## $ Longitude <dbl> -79.35139, -79.35139, -79.29333, -7…  
## $ EPT\_S <dbl> 16, 11, 29, 0, 30, 26, 4, 22, 20, 2…  
## $ BI <dbl> 4.354717, 4.891729, 4.505614, 8.085…  
## $ `% Impervious` <dbl> 0.005874221, 0.005874221, 0.0084387…  
## $ Drainage <dbl> 3.8100, 3.8100, 11.2000, 0.5700, 15…  
## $ Elevation <dbl> 319, 319, 258, 373, 271, 271, 384, …  
## $ Boulder <chr> "0", "0", "0", "0", "0", "0", "0", …  
## $ Cobble <chr> "0", "0", "0", "0", "0", "0", "0", …  
## $ Gravel <chr> "0", "0", "0", "0", "0", "5", "10",…  
## $ Sand <chr> "99", "99", "70", "30", "90", "85",…  
## $ Silt <chr> "0", "0", "20", "30", "10", "5", "1…  
## $ Other\_Substrate <chr> "<Null>", "<Null>", "10", "<Null>",…  
## $ Sub\_Desc <chr> NA, NA, "DETRITUS", NA, NA, NA, NA,…  
## $ Canopy <chr> "6", "1", "100", "100", "20", "70",…  
## $ Temp\_C <chr> "<Null>", "<Null>", "25.4", "<Null>…  
## $ Sp\_Cond <chr> "<Null>", "<Null>", "23.4", "<Null>…  
## $ pH\_SU <chr> "<Null>", "<Null>", "5.099999999999…  
## $ Diss\_Oxy <chr> "<Null>", "<Null>", "5.8", "<Null>"…  
## $ Stream\_Width <chr> "1", "1", "3", "1", "6", "4", "2", …  
## $ Channel\_Width <chr> "<Null>", "<Null>", "5", "<Null>", …  
## $ Avg\_Stream\_Depth <chr> "0.3", "0.2", "0.3", "0.2", "0.8", …  
## $ MaxStream\_Depth <chr> "<Null>", "<Null>", "0.8", "<Null>"…  
## $ Channel\_Modification <chr> "<Null>", "<Null>", "15", "<Null>",…  
## $ Instream\_Habitat <chr> "<Null>", "<Null>", "16", "<Null>",…  
## $ Substrate <chr> "<Null>", "<Null>", "7", "<Null>", …  
## $ Shade <chr> "<Null>", "<Null>", "10", "<Null>",…  
## $ Region <chr> "<Null>", "<Null>", "<Null>", "<Nul…  
## $ Basin <chr> "<Null>", "<Null>", "<Null>", "<Nul…  
## $ FID\_Haw\_Upper\_Nuese\_piedmont <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ FID\_Haw\_Upper\_Nuese <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ BASIN <chr> "Cape Fear", "Cape Fear", "Cape Fea…  
## $ FID\_Piedmont\_statecl <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…

### Data Cleaning and Formatting

The first water quality dataset was processed to specify the name of the creek or river and the watershed in which it runs. Measurement type was filtered to include only six types: fecal coliform, inorganic nitrogen, Kjeldahl (organic) nitrogen, specific conductance, total phosphorus, and turbidity, and unit conversions were done to ensure all measurements in each category had the same unit. Measurments noted as being below the analytical threshold for that technique were included as zero values. The data were also filtered by date so that all measurements included were taken during the growing season, which in North Carolina occurs between March and November.

The macroinvertibrate data were filtered to include only the measurements of biotic index along with geographic and temporal identification data. These data were then formatted to match the other ecological measurement data and the two datasets were merged. A small portion of this formatted data can be seen below.

## # A tibble: 116,061 x 8  
## Creek Month Day Year Measurement Value Unit WS   
## <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr> <chr>  
## 1 Deep 4 29 2013 Total Phosphorus 0.16 mg/L Haw   
## 2 Poplar 8 16 2001 Total Phosphorus 0.24 mg/L Neuse  
## 3 Rocky 3 3 2012 Kjeldahl Nitrogen 1.4 mg/L Haw   
## 4 Haw 4 12 1995 Inorganic Nitrogen 3.2 mg/L Haw   
## 5 Pigeon House 5 12 2009 Fecal Coliform 105 cfu/100… Neuse  
## 6 Neuse 6 20 2005 Kjeldahl Nitrogen 0.32 mg/L Neuse  
## 7 Middle 6 13 1997 Specific Conductan… 143 uS/cm Neuse  
## 8 Little 10 3 2012 Fecal Coliform 280 cfu/100… Haw   
## 9 Middle 10 20 2003 Turbidity 4.9 NTU Neuse  
## 10 Deep 10 6 2008 Turbidity 16 NTU Haw   
## # … with 116,051 more rows

In order to determine the relationships between each of the water quality variables, the data were organized and formatted by determining each instance where all six measurements were taken in the same location on the same day. These events are idenified by creek name, date, and coordinates of the testing site. Once these events are identified, the water quality dataset is filtered to include only data collected from them. These data are organized so that each row represents a measurement event with six columns, one for each nutrient measurement.

# Identify each instance where all six measurements were taken in a new column by concatenating the creek name, the date the measurement was taken, and the coordinates of the testing site.  
  
WQ\_wCDLL <- Water\_Quality %>%  
 mutate(CreekName = Creek,  
 Date = mdy(paste(Month, Day, Year, sep = "/")),  
 Date2 = Date, Lat2 = Lat, Long2 = Long) %>%  
 unite(col = "LatLong", Lat2, Long2, sep = ", ") %>%  
 unite(col = "Creek\_Date\_LatLong", CreekName, Date2, LatLong, sep = "\_")  
  
CDLL <- WQ\_wCDLL %>%  
 group\_by(Creek\_Date\_LatLong, Measurement) %>%  
 summarise(N = n()) %>%  
 filter(Measurement != "Biotic Index (BI)") %>%  
 spread(value = N, key = Measurement) %>%  
 rename("FC" = `Fecal Coliform`,  
 "IN" = `Inorganic Nitrogen`,  
 "KN" = `Kjeldahl Nitrogen`,  
 "TP" = `Total Phosphorus`,  
 "SC" = `Specific Conductance`,  
 "TU" = `Turbidity`) %>%  
 filter(FC+IN+KN+TP+SC+TU == 6) %>%  
 dplyr::select(Creek\_Date\_LatLong)

## `summarise()` regrouping output by 'Creek\_Date\_LatLong' (override with `.groups` argument)

# Filter data to include only instances where all six ecological markers were measured in the same location on the same day. Spread data so that each ecological marker is it's own column and each observation includes all six measurements. Also create total nitrogen (TN) column by adding Kjeldahl and Inorganic Nitrogen columns together.  
  
WQ <- WQ\_wCDLL %>%  
 filter(Creek\_Date\_LatLong %in% CDLL$Creek\_Date\_LatLong) %>%  
 arrange(Creek\_Date\_LatLong) %>%  
 dplyr::select(-Unit, -Month, -Day, -Year) %>%  
 spread(key = Measurement, value = Value) %>%  
 rename("FC" = `Fecal Coliform`,  
 "IN" = `Inorganic Nitrogen`,  
 "KN" = `Kjeldahl Nitrogen`,  
 "TP" = `Total Phosphorus`,  
 "SC" = `Specific Conductance`,  
 "TU" = `Turbidity`) %>%  
 mutate(TN = IN + KN) %>%  
 dplyr::select(Creek, Date, FC, SC, IN, KN, TN, TP, TU, Lat, Long)

### Analysis of Ecological Data

The nitrogen measurements in the dataset skew much higher than expected for typical waterways in the included watersheds. This is likely due to an overrepresentation of testing sites downstream of wastewater treatment plants. A K-means cluster analysis was performed to identify testing events that are likely effected by wastewater treatment plant discharge. The water quality dataset was filtered to exclude these suspected outliers.

# Cluster Analysis to exclude total nitrogen outliers likely caused by measurements downstream of waste-water treatment plants.  
  
# K-Means Cluster Analysis  
kWQ <- WQ$TN # Narrow dataset to total nitrogen measurements  
fit <- kmeans(kWQ, 2) # Create 2 clusters using K-means  
  
# Append cluster assignment to dataframe  
WQ\_Cluster <- data.frame(WQ, fit$cluster) %>%  
 filter(fit.cluster == 1)  
# Display mean total nitrogen values for each cluster group  
aggregate(kWQ,by=list(fit$cluster),FUN=mean)

## Group.1 x  
## 1 1 10.609163  
## 2 2 1.548702

The fecal coliform data featured several outliers with extremely high values (Z-value > 45) which have an outsize influence on summary statistics and on the sample data generation process. A similar k-means cluster analysis was performed, and suspected outliers were removed from the dataset.

# Cluster Analysis of fecal coliform to exclude outliers  
WQ\_Cluster %>%  
 dplyr::select(FC) %>%  
 summarise(Med = median(FC), # Summarize fecal coliform data quantiles to demonstrate extreme skew of the data.  
 Q75 = quantile(FC, 0.75),  
 Q90 = quantile(FC, 0.90),  
 Q95 = quantile(FC, 0.95),  
 Q99 = quantile(FC, 0.99),  
 Max = max(FC),  
 Avg = round(mean(FC), digits = 2))

## Med Q75 Q90 Q95 Q99 Max Avg  
## 1 160 400 1200 2684 11964 120000 808.49

fWQ <- WQ\_Cluster$FC  
fitFC <- kmeans(fWQ, 2)  
  
# Append cluster assignment to dataframe  
FC\_Cluster <- data.frame(WQ\_Cluster, fitFC$cluster)  
# Display mean fecal coliform values for each cluster  
aggregate(fWQ,by=list(fitFC$cluster),FUN=mean)

## Group.1 x  
## 1 1 86000.0000  
## 2 2 596.2201

Biotic index (BI) data are added to the dataset; however BI measurements were not made during the same measurement events and at the same sites as the nutrient values were collected. BI measurements are much more consistent for a particular stream than nutrient levels. Average BI measurements were calculated for each stream, and are joined to each row of nutrient data by stream.

# Filter biotic index data to exclude creeks from outside the target watersheds. Calculate average BI data for each creek.  
BI\_Join <- BI\_Data %>%  
 filter(Creek %in% unique(WQ\_Cluster$Creek)) %>%  
 group\_by(Creek) %>%  
 summarise(BI = round(mean(Value), digits = 2))

## `summarise()` ungrouping output (override with `.groups` argument)

# Join BI data to nutrient measurements such that each set of five nutrient measurements is paired with the average BI value for the particular stream it describes. (BI is much less variable than nutrient measurements for a given stream.)  
WQ\_Final <- WQ\_Cluster %>%  
 left\_join(BI\_Join, by = "Creek") %>%   
 dplyr::select(Creek, BI, FC, SC, TN, TP, TU) %>%  
 filter(FC < 7200) # Filter to remove fecal coliform outliers

### Sample Data Generation

Data are prepared for the generation algorithm by eliminating zero values so that the logorithm of the dataset can be taken, since the distribution is assumed to be log-normal. The logorithm of the dataset is taken to normalize the distribution, a sample dataset with 1000 rows is generated, and the data is converted back to the original log-normal distribution.

# Prepare data for data generation algorithm. Zero values for nutrient measurements will create NA values when the logorithm is taken. The minimum non-zero value for each measurement is taken as the minimum detection threshold, and all zero values are replaced with this value.  
  
WQ\_Final$FC[WQ\_Final$FC == 0] <- 1  
WQ\_Final$TP[WQ\_Final$TP == 0] <- 0.01  
WQ\_Final$TU[WQ\_Final$TU == 0] <- 0.1  
  
# Create vectors and matrices to enable sample data generation algorithm. Normalize lognormal data before generating sample data - then return to lognormal format  
  
Column\_Means <- as.numeric(colMeans(WQ\_Final[,2:7], na.rm = TRUE))  
  
LogNorm <- WQ\_Final[,2:7]  
Norm <- log(LogNorm)  
  
NormCov <- cov(Norm, use = "pairwise.complete.obs")  
NormMeans <- log(Column\_Means)  
  
NormSample <- mvrnorm(n = 1000, mu = NormMeans, Sigma = NormCov)  
LogNormSample <- exp(NormSample)  
Sample\_Data <- round(LogNormSample, 3)

## # A tibble: 5 x 6  
## BI FC SC TN TP TU  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 5.44 6363. 787. 5.44 5.66 6.38  
## 2 6.27 489. 987. 10.2 1.31 2.80  
## 3 5.58 49.9 1016. 18.1 0.739 18.0   
## 4 6.44 785. 561. 11.5 1.92 1.46  
## 5 5.07 449. 416. 5.13 1.45 9.69

### Selection of Sample Data Rows

Z-Values are calculated for each ecological data point in the generated sample data. For each variable, the 50 data rows with greatest magnitude Z-value are selected. These rows are narrowed by excluding any row which features in the top 50 Z-values for more than one variable, and then 12 rows are selected at random for each variable, making up the first 72 data rows.

T50 <- function(ColNum, ColName, Meas) {  
 tibble(Row = ZVal[[1]],  
 Z = abs(ZVal[[ColNum]])) %>%  
 arrange(desc(Z)) %>%  
 head(n = 50) %>%  
 mutate(Meas = Meas)  
}

UniqueRows <- Top50 %>%  
 group\_by(Row) %>%  
 summarise(N = n()) %>%  
 filter(N == 1)

## `summarise()` ungrouping output (override with `.groups` argument)

RowSelect <- Top50 %>%  
 filter(Row %in% UniqueRows$Row) %>%  
 group\_by(Meas) %>%  
 sample\_n(12)

The remaining rows are selected by filtering the data to include only the lowest 80 percent of measurements for each water quality variable, and randomly selecting 40 observations from this data set. The 72 data rows representing the extremes of each variable and the 40 data rows that represent the more normal range of each variable are combined to form the final set of 112 generated data rows.

Pct80 <- Sample\_Data %>%  
 gather(value = Value, key = Measurement) %>%  
 group\_by(Measurement) %>%  
 summarise(Pct80 = quantile(Value, 0.8)) %>%  
 dplyr::select(Pct80) %>%  
 t() %>% # Determine the 80th percentile for each variable;  
 as.numeric() # store these values in a vector: Pct80

## `summarise()` ungrouping output (override with `.groups` argument)

Sample100 <- rbind(  
 Sample\_Data %>% # Filter sample data to include selected rows from  
 mutate(Row = 1:1000) %>% # top five percent of Z-values for each variable  
 filter(Row %in% RowSelect$Row) %>%  
 left\_join(RowSelect, by = "Row") %>%  
 dplyr::select(BI, FC, SC, TN, TP, TU, Meas, Z),  
   
 Sample\_Data %>% # Filter the sample data to exclude the top  
 filter(BI < Pct80[1], # 20 percent for each variable.  
 FC < Pct80[2],  
 SC < Pct80[3],  
 TN < Pct80[4],  
 TP < Pct80[5],  
 TU < Pct80[6]) %>%  
 sample\_n(40) %>%  
 mutate(Meas = NA, Z = NA)  
 ) %>%  
 arrange(Meas, desc(Z))  
  
colnames(Sample100) <- c("Biotic Index", "Fecal Coliform (cfu/100mL)",   
 "Specific Conductance (uS/cm)",   
 "Total Nitrogen (mg/L)", "Total Phosphorus (mg/L)",   
 "Turbidity (NTU)", "Key Variable",   
 "Z-Value of Key Variable")  
  
glimpse(Sample100)

## Rows: 112  
## Columns: 8  
## $ `Biotic Index` <dbl> 7.901, 7.590, 4.347, 4.359, 7.543…  
## $ `Fecal Coliform (cfu/100mL)` <dbl> 245.035, 1291.660, 211.966, 1463.…  
## $ `Specific Conductance (uS/cm)` <dbl> 350.753, 520.634, 239.215, 612.24…  
## $ `Total Nitrogen (mg/L)` <dbl> 13.004, 12.323, 4.552, 7.867, 13.…  
## $ `Total Phosphorus (mg/L)` <dbl> 0.482, 0.416, 1.329, 0.769, 0.959…  
## $ `Turbidity (NTU)` <dbl> 28.207, 7.264, 24.969, 5.254, 22.…  
## $ `Key Variable` <chr> "BI", "BI", "BI", "BI", "BI", "BI…  
## $ `Z-Value of Key Variable` <dbl> 3.008291, 2.524403, 2.521412, 2.5…

The “Key Variable” column indicates which variable in the data row was among the top 50 in terms of the magnitude of the Z-value for that variable. Rows that were selected from the bottom 80 percent portion do not have a key variable. This is for visual organization and inspection of the data, and is not indicated to experts completing the elicitation.