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.

## Observations: 6  
## Variables: 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.

## Observations: 1,234  
## Variables: 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 11 3 2009 Turbidity 8.4 NTU Haw   
## 2 Middle 8 25 2005 Specific Conductan… 174 uS/cm Neuse  
## 3 Muddy 6 27 2005 Fecal Coliform 230 cfu/100… Haw   
## 4 Muddy 10 2 2003 Fecal Coliform 460 cfu/100… Haw   
## 5 Haw 8 30 2006 Total Phosphorus 0.84 mg/L Haw   
## 6 Northeast 4 24 2013 Inorganic Nitrogen 0.16 mg/L Haw   
## 7 Little 9 17 2008 Inorganic Nitrogen 0.15 mg/L Haw   
## 8 Big Governo… 10 3 2013 Fecal Coliform 82 cfu/100… Haw   
## 9 Deep 5 17 1995 Turbidity 5.3 NTU Haw   
## 10 Neuse 5 16 2001 Specific Conductan… 113 uS/cm Neuse  
## # … with 116,051 more rows

Text

# Determine each instance where inorganic and kjeldahl nitrogen were measured in the same river, at the same location, and on the same day. Identify each instance in a new column by concatenating the creek name, the date the measurement was taken, and the coordinates of the testing site.  
  
Nitro\_Sample\_Sites <- Water\_Quality %>%  
 filter(Measurement %in% c("Kjeldahl Nitrogen", "Inorganic Nitrogen")) %>%  
 mutate(Date = mdy(paste(Month, Day, Year, sep = "/"))) %>%  
 unite(col = "LatLong", Lat, Long, sep = ", ") %>%  
 group\_by(Date, Creek, Measurement, LatLong) %>%  
 summarise(N = n()) %>%  
 spread(key = Measurement, value = N) %>%  
 filter(`Inorganic Nitrogen` == 1 & `Kjeldahl Nitrogen` == 1) %>%  
 dplyr::select(Date, Creek, LatLong) %>%   
 tidyr::unite(col = "Creek\_Date\_LatLong", Creek, Date, LatLong, sep = "\_")   
  
# Filters the general dataset to include only sets of kjeldahl and inorganic nitrogen taken in the same place at the same time. Total nitrogen can be calculated by adding inorganic and Kjeldahl (organic) nitrogen when the two measurements are taken together.  
  
Total\_Nitrogen <- Water\_Quality %>%  
 filter(Measurement %in% c("Kjeldahl Nitrogen", "Inorganic Nitrogen")) %>%  
 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 = "\_") %>%  
 filter(Creek\_Date\_LatLong %in% Nitro\_Sample\_Sites$Creek\_Date\_LatLong) %>%  
 arrange(Date, Creek) %>%  
 spread(key = Measurement, value = Value) %>%  
 mutate(Value = `Inorganic Nitrogen` + `Kjeldahl Nitrogen`, Measurement = "Total Nitrogen") %>%  
 dplyr::select(-Creek\_Date\_LatLong, -`Inorganic Nitrogen`, -`Kjeldahl Nitrogen`, -Date)   
  
# Replace inorganic and Kjeldahl nitrogen with the total nitrogen measurements.  
  
Water\_Quality\_wTN <- Water\_Quality %>%  
 filter(!Measurement %in% c("Kjeldahl Nitrogen", "Inorganic Nitrogen")) %>%  
 rbind(Total\_Nitrogen)

# Calculate average values for each ecological measurement in each creek  
  
CreekMeans <- Water\_Quality\_wTN %>%  
 group\_by(Creek, Measurement) %>%  
 summarise(Avg = mean(Value)) %>%  
 spread(key = Measurement, value = Avg)  
  
# Create a vector of average values for each measurement  
  
SampleMeans <- Water\_Quality\_wTN %>%  
 group\_by(Measurement) %>%  
 summarise(Avg = mean(Value)) %>%  
 dplyr::select(Avg) %>%  
 t() %>%  
 as.numeric()  
  
# Create a vector of standard deviations for each measurement  
  
SampleSDs <- Water\_Quality\_wTN %>%  
 group\_by(Measurement) %>%  
 summarise(SD = sd(Value)) %>%  
 dplyr::select(SD) %>%  
 t() %>%  
 as.numeric()

# Normalize lognormal data before generating sample data - then return sample to lognormal distribution.  
  
LogNorm <- CreekMeans[,2:7]  
Norm <- log(LogNorm)  
  
NormCov <- cov(Norm, use = "pairwise.complete.obs")  
NormMeans <- log(SampleMeans)  
  
NormSample <- mvrnorm(n = 1000, mu = NormMeans, Sigma = NormCov)  
LogNormSample <- exp(NormSample)  
SampleData <- round(LogNormSample, 3)  
  
colnames(SampleData) <- c("Biotic Index (BI)", "Fecal Coliform (cfu/100mL)", "Phosphorus (mg/L)",   
 "Specific Conductance (uS/cm)", "Total Nitrogen (mg/L)", "Turbidity (NTU)")  
  
head(SampleData, n = 5)

## Biotic Index (BI) Fecal Coliform (cfu/100mL) Phosphorus (mg/L)  
## [1,] 5.461 896.865 223.337  
## [2,] 4.695 1140.389 364.378  
## [3,] 7.333 856.090 262.101  
## [4,] 5.165 3641.836 174.413  
## [5,] 4.719 121.064 136.650  
## Specific Conductance (uS/cm) Total Nitrogen (mg/L) Turbidity (NTU)  
## [1,] 4.226 0.711 15.457  
## [2,] 1.704 0.168 21.505  
## [3,] 1.122 0.073 9.354  
## [4,] 2.752 0.180 40.967  
## [5,] 0.898 0.038 12.343

# Create a dataframe of the Z-value for each datum in the generated sample data.  
  
ZValue <- function(Data, Means, SDs) {  
 Z1 = (Data[,1] - Means[1])/SDs[1]  
 Z2 = (Data[,2] - Means[2])/SDs[2]  
 Z3 = (Data[,3] - Means[3])/SDs[3]  
 Z4 = (Data[,4] - Means[4])/SDs[4]  
 Z5 = (Data[,5] - Means[5])/SDs[5]  
 Z6 = (Data[,6] - Means[6])/SDs[6]  
   
 return(cbind(Z1, Z2, Z3, Z4, Z5, Z6, c(1:1000)))  
}  
  
ZVal <- ZValue(Data = SampleData, Means = SampleMeans, SDs = SampleSDs)

T50 <- function(ColNum, ColName, Meas) {  
 tibble(Row = ZVal[,7],   
 Z = abs(ZVal[,ColNum])) %>%  
 arrange(desc(Z)) %>%  
 head(n = 50) %>%  
 mutate(Meas = Meas)  
}  
  
Top50 <- rbind(  
T50(ColNum = 1, ColName = BI, Meas = "BI"),  
T50(ColNum = 2, ColName = FC, Meas = "FC"),  
T50(ColNum = 3, ColName = SC, Meas = "SC"),  
T50(ColNum = 4, ColName = TN, Meas = "TN"),  
T50(ColNum = 5, ColName = TP, Meas = "TP"),  
T50(ColNum = 6, ColName = TU, Meas = "TU"))  
   
UniqueRows <- Top50 %>%  
 group\_by(Row) %>%  
 summarise(N = n()) %>%  
 filter(N == 1)  
  
RowSelect <- Top50 %>%  
 filter(Row %in% UniqueRows$Row) %>%  
 group\_by(Meas) %>%  
 sample\_n(12)

Pct80 <- SampleData %>%  
 gather(value = Value, key = Measurement) %>%  
 group\_by(Measurement) %>%  
 summarise(Pct80 = quantile(Value, 0.8)) %>%  
 dplyr::select(Pct80) %>%  
 t() %>%  
 as.numeric()  
  
Sample100 <- rbind(  
 SampleData %>% # 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),  
   
 SampleData %>% # 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")  
  
print(Sample100)

## # A tibble: 112 x 8  
## `Biotic Index` `Fecal Coliform… `Specific Condu… `Total Nitrogen…  
## <dbl> <dbl> <dbl> <dbl>  
## 1 10.6 1733. 174. 2.40   
## 2 8.11 119. 216. 2.22   
## 3 8.04 592. 326. 2.42   
## 4 7.98 1369. 226. 2.05   
## 5 7.94 1181. 187. 3.58   
## 6 7.88 560. 164. 0.848  
## 7 7.84 609. 227. 2.92   
## 8 3.54 213. 144. 1.12   
## 9 7.69 271. 174. 2.12   
## 10 3.66 344. 185. 2.06   
## # … with 102 more rows, and 4 more variables: `Total Phosphorus  
## # (mg/L)` <dbl>, `Turbidity (NTU)` <dbl>, `Key Variable` <chr>, `Z-Value  
## # of Key Variable` <dbl>