# Sediment Core Analysis

*Rebecca Stubbs*

*February 16, 2018*

## Prompt

Background: The attached rds file (Dataset_B) contains chemistry data for sediment cores from an urban waterway (downloaded from here and slightly modified to achieve the attached format). This urban waterway has a complex history of discharges from storm water, sanitary sewer, and industrial sources as well as operations that took place up to 50 years prior to when samples were collected. Over time, substantial anthropogenic modifications (e.g. dredging and disturbance) have occurred. The key variables of interest within this dataset include:

* Coordinates: X, Y
* Chemical names: Chemical
* Sample date: SampleDate
* Analytical chemical results: ValueOrHalfQL
* Result units: Unit
* Detection status (was the result above or below the analytical method detection limit): Detected
* Flag indicating if the sample is in the dredged navigation channel: FE_NAVCHAN_F
* Core segment upper/lower depth: UpperDepth_ft , LowerDepth_ft, UpperDepth_cm, LowerDepth_cm
* Elevation of sediment at coordinate where sample was collected: FE_Z2003Bathy_MLLW

Given the attached dataset:

1. Divide the site into spatial and depth "zones" (groups, regions, strata) based on chemistry of Total PCBs, Arsenic, and PAHs.
2. Are there upstream to downstream patterns? If so, what are the differences between chemicals?
3. Describe the trends you found and the tools/approaches that you used. Provide PDF maps, tables, and statistics depicting the spatial zones, strata, and groupings.
4. Are there any additional useful interpretations that you can identify?

Bonus: Can you infer anything about the modification history of the site from the chemistry data? If so, what did you find? What tools, approach, algorithms, and statistics did you use?

# Analysis

## General Thought Process

This data represents sediment cores taken along the last ~4 miles of the Duwamish river, in Washington State– not so far from where I lived for the last few years. There is a *lot* going on here– between the extended time period of the samples, to core depth, dredging history, and chemical detection, this is a hugely multi-dimensional data set before you even start exploring the X and Y (much less the Z of the bathymetric river depth).

A full and satisfying answer to these questions could be a major undertaking– to get a complete picture of what's happening in the river bed, at different depths, and over time, it would be helpful to "fill in the gaps" between observations, essentially creating a 3-d mesh of estimated chemical concentrations over time, at which point I could generate thresholds that define different levels of contamination requiring different intervention strategies or indicating historical activity. To accomplish this would most likely involve a complex geostatistical model, in which spatial and temporal auto-correlation, information about previous dredging, river depth, proximity to known or suspected contamination sites, and potentially values from other nearby chemicals, are used to estimate each analyte for any given space, time, and sediment depth. Incorporating historical flooding and rainfall information might also help understand downstream sediment patterns. Using polygon layers of previous dredging efforts to provide "dredged or undredged" information for spaces not observed by cores, a mesh of points could have their values "predicted" by this geostatistical model, at which point, the river could be divided based on either individual contaminant values and thresholds, or based on the combinations of total PCBs, PAHs, and arsenic. The extent to which "lags", or observations prior (either upstream, or back in time) helped predict each of the chemicals within the model could potentially indicate the presence of upstream-to-downstream contaminant flow.

Regretfully, I don't currently have the time to tackle such an analysis– furthermore, the "less intense" version of this process (simply interpolating between values across river miles and/or time) is displeasing in a number of ways, and would require a vast array of assumptions that may not accurately represent the processes at work within the Duwamish. Interpolation involves stretching values across a data gap between 2 points, with values stretching in-between. However, for circumstances where there are gaps between data points that are low, then high, then low as you travel downstream, it would be counter-intuitive to stretch the values between the low value upstream, to the high value upstream, such that pollution increases incrementally as you go downriver.

I have avoided interpolating between observed values, and have divided the site into four sediment depth zones: surface, near-surface, mid-layer, and deep sediments. To keep the focus of this analysis on the relatively undisturbed sediments without regular dredging, I consider only samples that are not taken within the dredged shipping channel. Samples taken in other dredged areas (the data points for which there is a valid DredgeYear, but not a flag for the data point being in the shipping channel) are included, since the sediment samples taken are a representation of what is "there"– it just might not reflect accurately the history of pollution deposition, which isn't the focus of this analysis.

Since the focus of this analysis is discovering upstream and downstream patterns, and not which bank of the river has more contaminants, I have reduced the spatial dimensionality from X, Y coordinates, down to 1 number that describes where, in terms of stream flow, each observation has occurred, based on river mile (distance away from the mouth of the river).

What follows are my explorations of the data set, which should start to probe around some of the trends present in PCBs, PAH, and arsenic in the lower Duwamish. There are many stones unturned in this investigation–with time as a limited resource, the following describes the first steps I would take to try to understand the spatio-temporal trends in contaminants.

**NB: This document is as a .pdf rendering of an R Markdown, which allows for the integration of documentation, code, and outputs. What follows is all of the code required to process the data, and generate the outputs for this exercise.**

## Data Prep

```
# Duwamish River sediment core analysis
# R Stubbs Feb 15 2016
# Input: .RDS of chemical core data in Duwamish R.
rm(list=ls()) # Clear working environment
library("MapSuite") #Self-written library, has many common libs as dependencies

# Load in data, which is a spatialpointsdataframe
  chem_sp<-readRDS("Dataset_B.RDS")

# Let's get the data out from the spatial object
  chem_data<-data.table(chem_sp@data)
  chem_data[,sp_index:=seq(1:nrow(chem_data))]

# OK, there are way too many unhelpful columns in this dataset; let's parse it down
  chem_data<-chem_data[,list(sp_index,X,Y,RM,LocationName,
                             UpperDepth_cm,LowerDepth_cm, Unit,
                             Chemical,ChemicalGroup, SampleDate,
                             Detected, observed_value=ValueOrHalfQL,
                             bathy=FE_Z2003Bathy_MLLW,
                             Dredged, DredgeYear, shipping_channel=FE_NAVCHAN_F)]

# make one field that describes what chemical we are interested in; that way we don't
# have to subset off of Chemical *or* ChemicalGroup depending on the analyte
  chem_data[Chemical=="Arsenic",chem:="Arsenic"]
  chem_data[Chemical=="Total PCBs",chem:="PCBs"]
  chem_data[ChemicalGroup=="PAHs",chem:="PAHs"]

# Make decimal date
  chem_data[,decimal_date:=lubridate::decimal_date(SampleDate)]

# Some basic cleanup: I only want to know about samples that have a valid entry
# for "Detected", and that have non-NA values for observed chemical value
  chem_data<-chem_data[!is.na(Detected)&!is.na(observed_value)]

# What year was each sample taken? Extract from the sample time.
  chem_data[,Year:=as.numeric(substr(as.character(chem_data$SampleDate),1,4))]

# Let's try a log-transform for the value variables; this might be useful later on
  chem_data[,log_value:=log(observed_value)]

# Adding a column of "noise" to be able to jitter our plots
  chem_data[,noise:=rnorm(nrow(chem_data),mean=0,sd=.2)]
```

To make looping through the different chemicals easier, I've made a named list with a specific color pallete to keep them straight, and their units.

```
# Making a structured list for each of the chemicals with data sets, units, and
chemlist<-list()
chemlist[["PAHs"]]<-list(colors="sky",name="PAHs",unit="ug/kg dw")
chemlist[["Arsenic"]]<-list(colors="ocean",name="Arsenic",unit="mg/kg dw")
chemlist[["PCBs"]]<-list(colors="berries",name="PCBs",unit="ug/kg dw")
```

I also find the fact that River-Miles are based on distance from the mouth of the stream's outlet to the ocean, rather than distance from the furthest upstream point, to be confusing, and I like my graph and chart axes such that "trajectory" of the river is moving towards the right in accordance with what is downstream. To make this more straightforward, I'll create a "flipped" river-mile counter, in which the measurement represents the distance from the furthest-upstream-point measured in the data set.

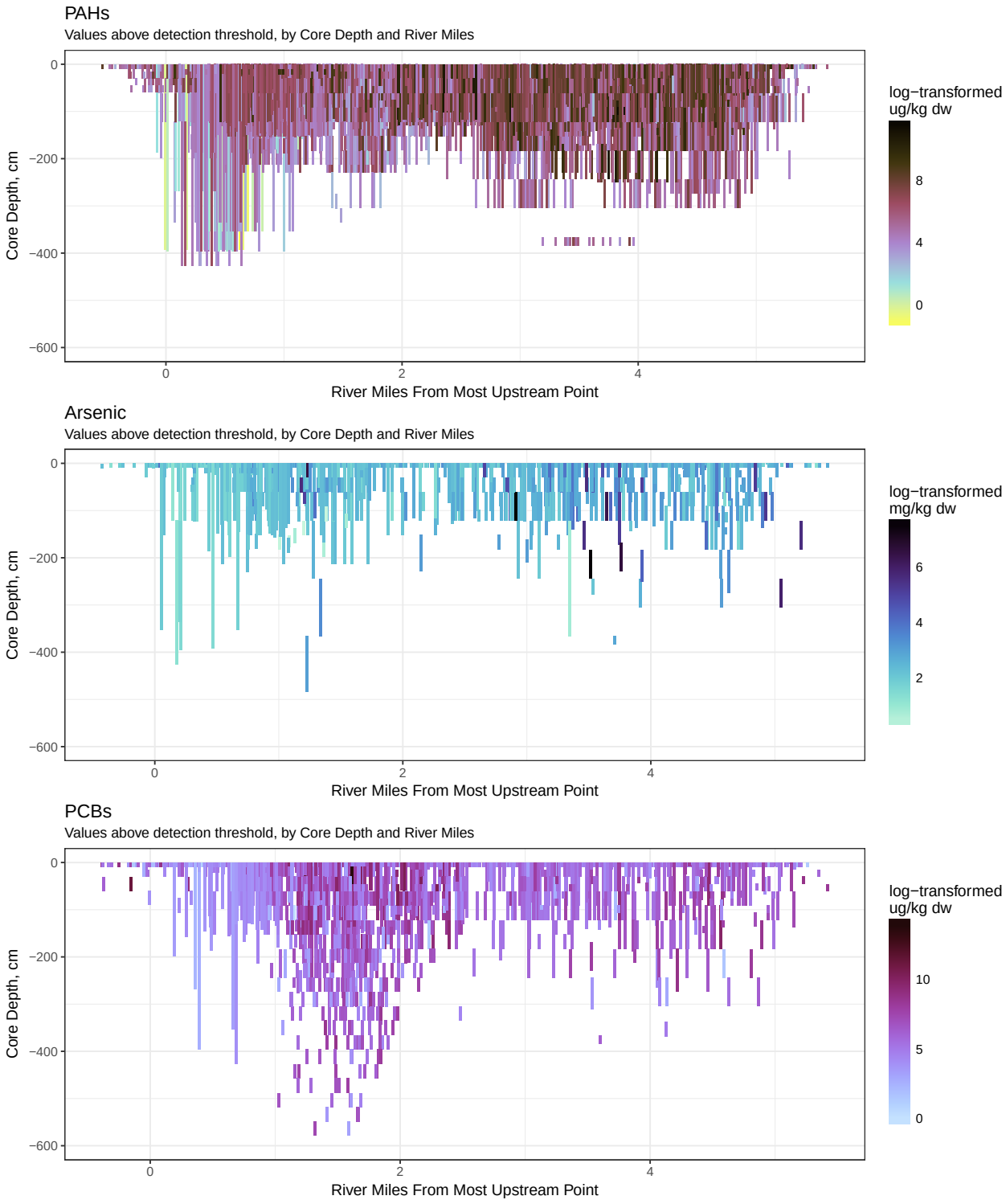
```
chem_data[,RM_up:=max(RM,na.rm=T)-RM]
```

## Defining sedimentary depth strata based on common sampling depths

The first step here is to determine different "layers" of sediment might be meaningful– to do this, I need to know what samples were taken, and where– is there a consistent pattern of depths cored, or a reasonable cut-point based on observed values within the depths sampled? To scope this out, I made some graphs by river-mile and depth core, to see what patterns appear in the data.

In these plots, the size of the bar on the Y axis represents the sedimentary cross-section for the data sample, while the X axis describes the distance away from the furthest measured point, in river miles. The color of the vertical bars reflects the log-transformed amount of chemical present in the sample, to highlight differences in the samples at low values.

```
for (chemical in names(chemlist)){ # For each chem
  chem_info<-chemlist[[chemical]] # get color/unit info

  # Make GGplot
  p1<-ggplot() +
    # Rectangles to simulate core depths and values
     geom_rect(data=chem_data[chem==chemical & Detected=="Yes",],
               # River-mile as x-axis, plus some noise for
               # differentian of samples and artificial width
                 aes(xmin=RM_up+noise, xmax=RM_up+noise+.02,
               # Use upper/lower limit as core sample depth limits
                 ymin=-1*UpperDepth_cm, ymax=-1*LowerDepth_cm,
                 fill=log(observed_value) )) + # Color by log-value

    # Formatting and Labeling
    ggtitle(chem_info[["name"]],
    subtitle="Values above detection threshold, by Core Depth and River Miles") +
    xlab("River Miles From Most Upstream Point") + ylab("Core Depth, cm") +
    theme_bw() + scale_fill_gradientn(colors=wpal(chem_info[["colors"]])) +
    guides(fill=guide_colourbar(title=paste0("log-transformed \n",chem_info[["unit"]]),
                                title.position="top", barheight=10,
                                barwidth=1, label=TRUE, ticks=FALSE,
                                direction="vertical"))+ylim(-600,0)
  print(p1)
  }
```



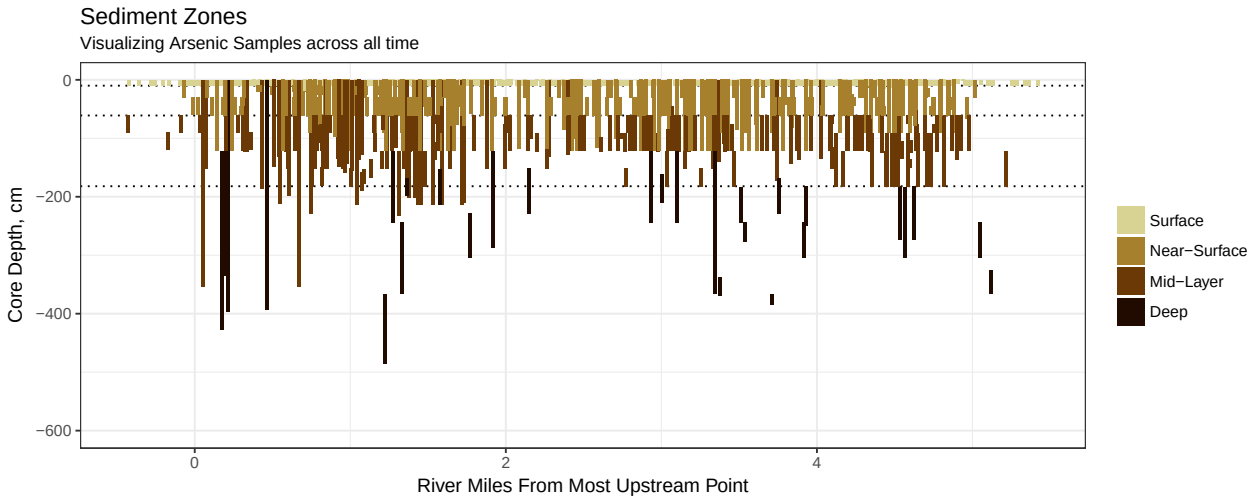

It doesn't appear that there is a deep core sampling depth that reflects lower values, as might be indicative of having measured the "background" levels of the contaminants. This isn't altogether unsurprising– pollution has been happening in this river for a long time before the sampling process began.

It appears that there are many different common core sampling depths, but, to keep this analysis tractable, we'll stick with 4 zones: "Surface", with the midpoint of the sample falling between 0-10 cm, "Near-Surface", with the midpoint depth of the sample falling between 10cm and 2 feet deep, "Mid-layer", with samples between 2-6 feet deep, and "Deep", with all samples beyond 6 feet below the river bottom. Some of these cores may span more than one of these categories– to avoid "double-counting" the data points in multiple sedimentary strata, I'll use the depth midpoint of the core sample to assign categories.

```
## Generate factor-data-type classification based
## on sample depth with defined cut-points
# Generate mid-point of core samples
chem_data[,sample_midpoint:=(LowerDepth_cm+UpperDepth_cm)/2]
chem_data[,Depth:=cut(sample_midpoint,breaks=c(0,10,61,182,10000),
           labels=c("Surface","Near-Surface","Mid-Layer","Deep"))]
chem_data[,Depth:=as.factor(Depth)]
```

Let's create the categories, and sanity-check that they seem to be in the right categories based on one of the chemicals:

```
# Define a color pallette for the factor variable, Depth
SedColors<-wpal("warm_brown",noblack=T,n=4)
names(SedColors) <- c("Surface","Near-Surface","Mid-Layer","Deep")

# Make GGplot
  zone_plot<-ggplot() +
  # Show zones as dotted lines underneath data plots

  # Boundary for Shallow and Near-Surface
    geom_hline(yintercept = -10, linetype="dotted") +
  # Boundary between Near-Surface and Mid-layer
    geom_hline(yintercept=-61, linetype="dotted") +
  # Boundary between Middle and Deep Zone
    geom_hline(yintercept=-182, linetype="dotted") +

  # Show cores in their respective zones
    geom_rect(data=chem_data[chem=="Arsenic",],
      aes(xmin=RM_up+noise, xmax=RM_up+noise+.02,
      ymin=-1*UpperDepth_cm, ymax=-1*LowerDepth_cm, fill=Depth)) +
    scale_fill_manual(name = "",values = SedColors, drop=F) +
    xlab("River Miles From Most Upstream Point") +
    ylab("Core Depth, cm") +ylim(-600,0) +
    ggtitle("Sediment Zones",
    subtitle="Visualizing Arsenic Samples across all time") + theme_bw()

 print(zone_plot)
```



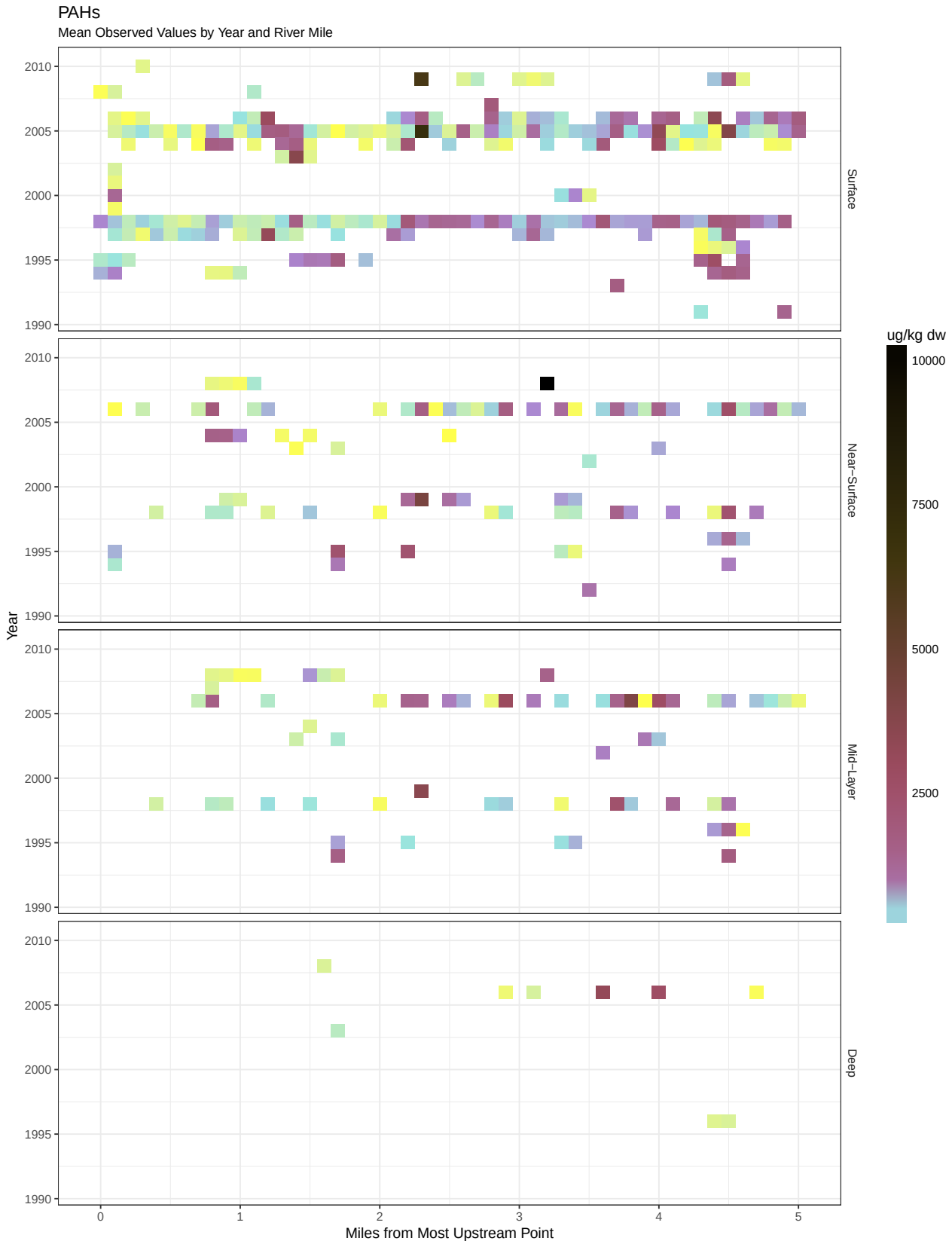

We can see that some of the samples do indeed cross strata lines, but for the most part, this looks good, and a relatively clean representation of depth strata for the samples.

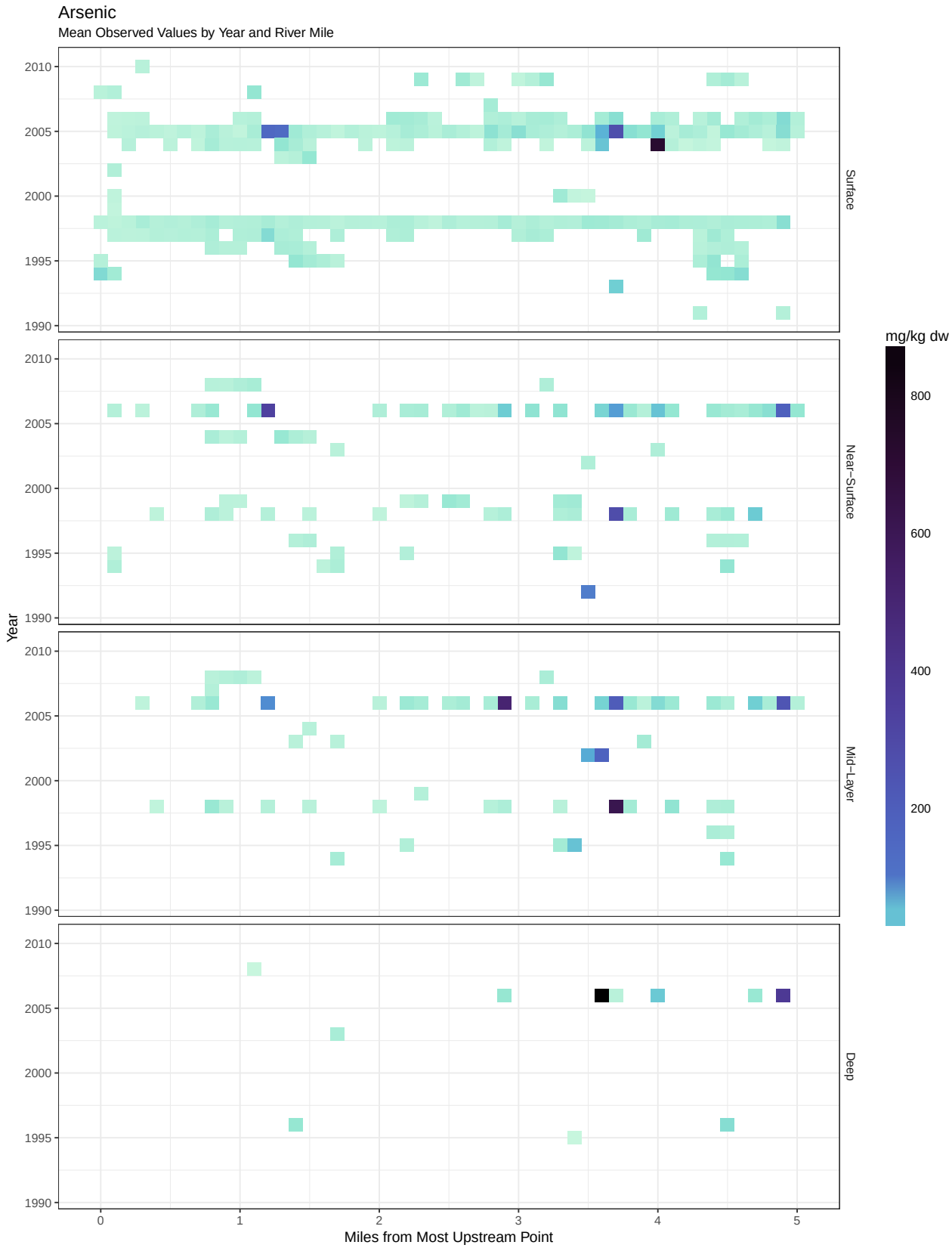
Since there are observations in each river-mile from one bank to the other, let's reduce the dimensionality even more, and create some summary statistics for each tenth of a river mile, and year combination.

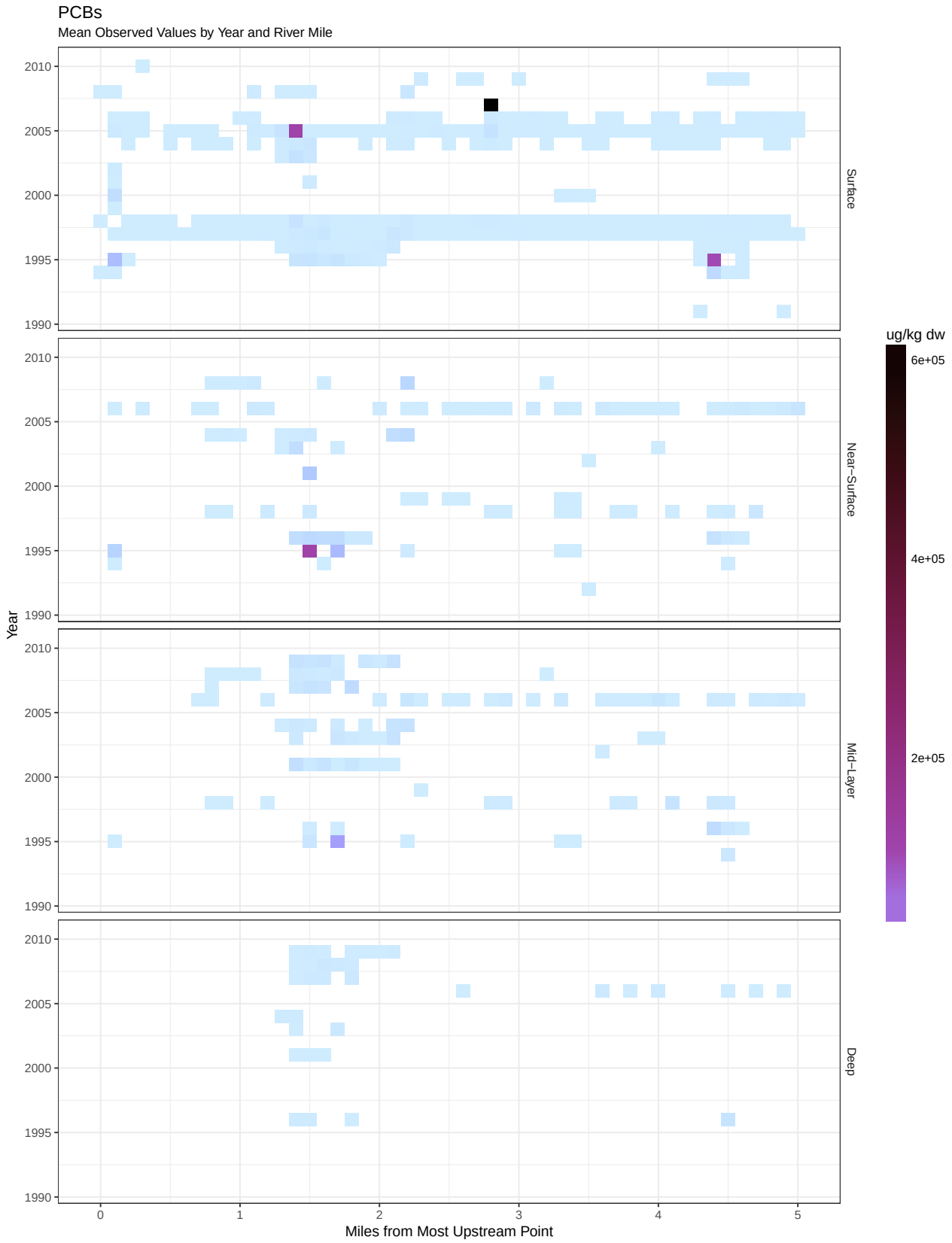
```
# Calc. mean and SD for each chemical, in each
# year, dredge-category, depth-category,river-tenth-mile
chem_simplified<-chem_data[!is.na(chem) & Detected=="Yes" &
                             shipping_channel=="N",
                             list(obs_mean=mean(observed_value),
                             obs_sd=sd(observed_value), Unit),
                             by=c("chem","Year","RM_up","shipping_channel","Depth")]
```

Now, I generate plots for each year, river-tenth-mile, and depth zone, to look for sweeping patterns across space, time, or sediment depth.

```
for(chemical in names(chemlist)){
  chem_info<-chemlist[[chemical]]

depth_plot<- ggplot() +
   ggtitle(chemical, subtitle="Mean Observed Values by Year and River Mile") +
   geom_tile(data=chem_simplified[chem==chemical,],
             aes(x=RM_up,y=Year, fill=(obs_mean))) + # Color by log-value
   xlab("Miles from Most Upstream Point") +
   theme_bw() + scale_fill_gradientn(colors=wpal(chem_info[["colors"]]), values=c(0,.1,1)) +
   guides(fill=guide_colourbar(title=paste0("",chem_info[["unit"]]),
                               title.position="top", barheight=30,
                               barwidth=1, label=TRUE, ticks=FALSE, direction="vertical")) +
  facet_wrap(~Depth,ncol=1,strip.position="right")+theme(strip.background = element_blank())
print(depth_plot)

}
```


PAHs
Mean Observed Values by Year and River Mile
Year
2010
2005
2000
1995
1990
Surface
Near-Surface
Mid-Layer
Deep
0
1
2
3
4
5
Miles from Most Upstream Point
ug/kg dw
10000
7500
5000
2500


Arsenic
Mean Observed Values by Year and River Mile
Year
2010
2005
2000
1995
1990
Surface
Near−Surface
Mid−Layer
Deep
0
1
2
3
4
5
Miles from Most Upstream Point
mg/kg dw
800
600
400
200


PCBs
Mean Observed Values by Year and River Mile
Year
2010
2005
2000
1995
1990
Surface
Near−Surface
Mid−Layer
Deep
0
1
2
3
4
5
Miles from Most Upstream Point
ug/kg dw
6e+05
4e+05
2e+05

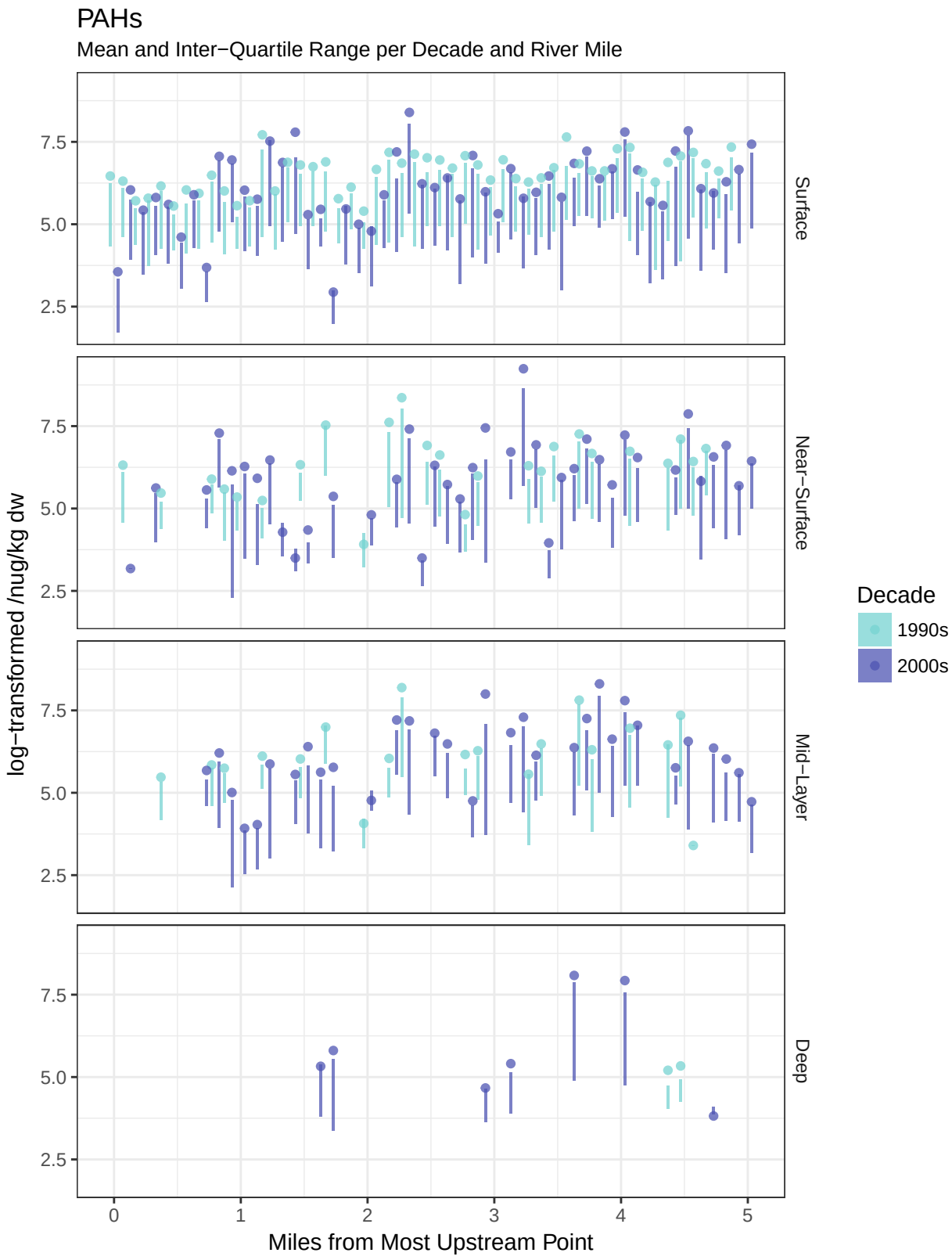
The above plots show very little about the variation of observations– to get a better sense of the distribution, the follow plots highlight not just the mean, but also the inter-quartile range of the data within each decade of observation, by river-mile.

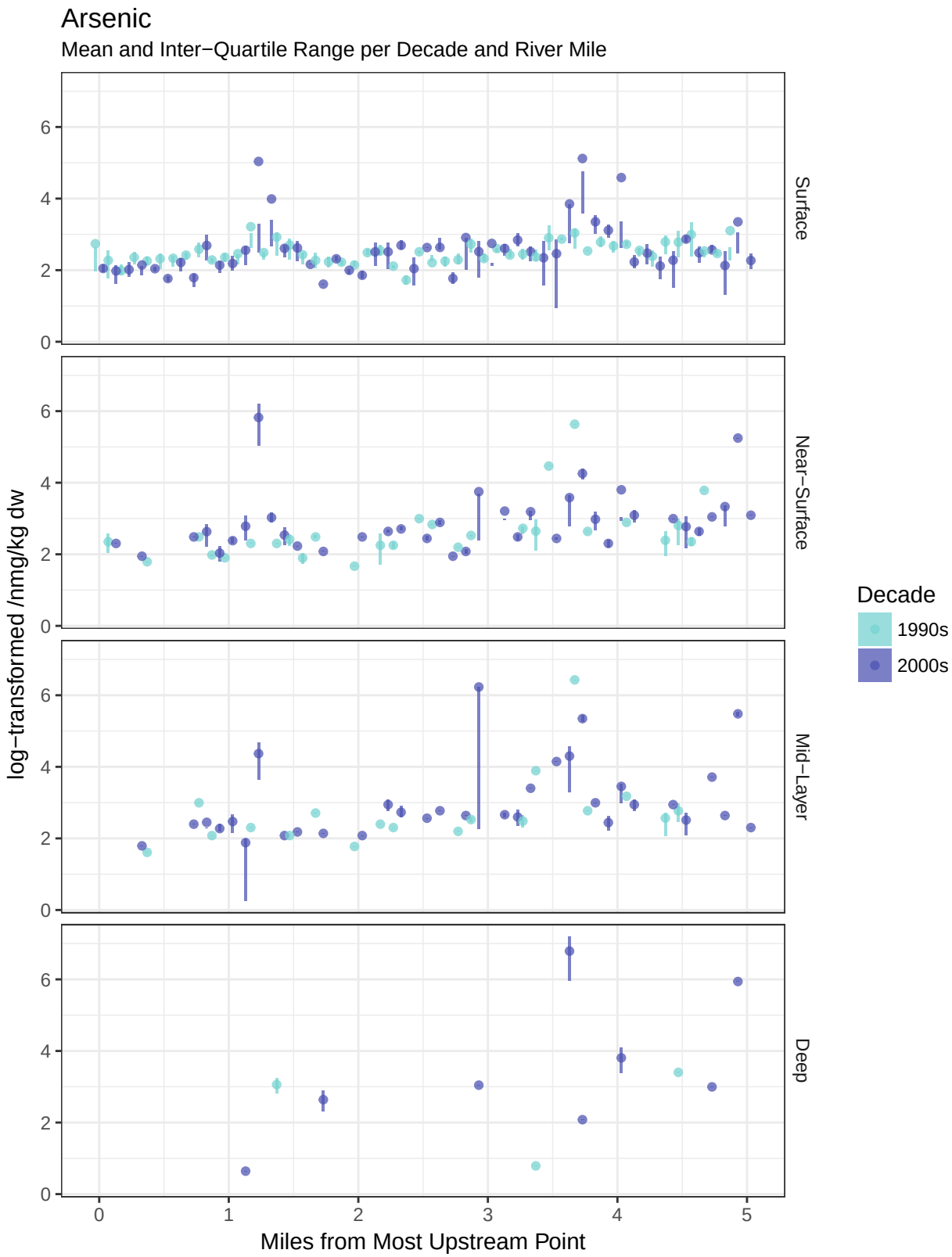
First, I calculate a new data set that summarizes by decade, calculating the mean and inter-quartile range of the data within each decade.

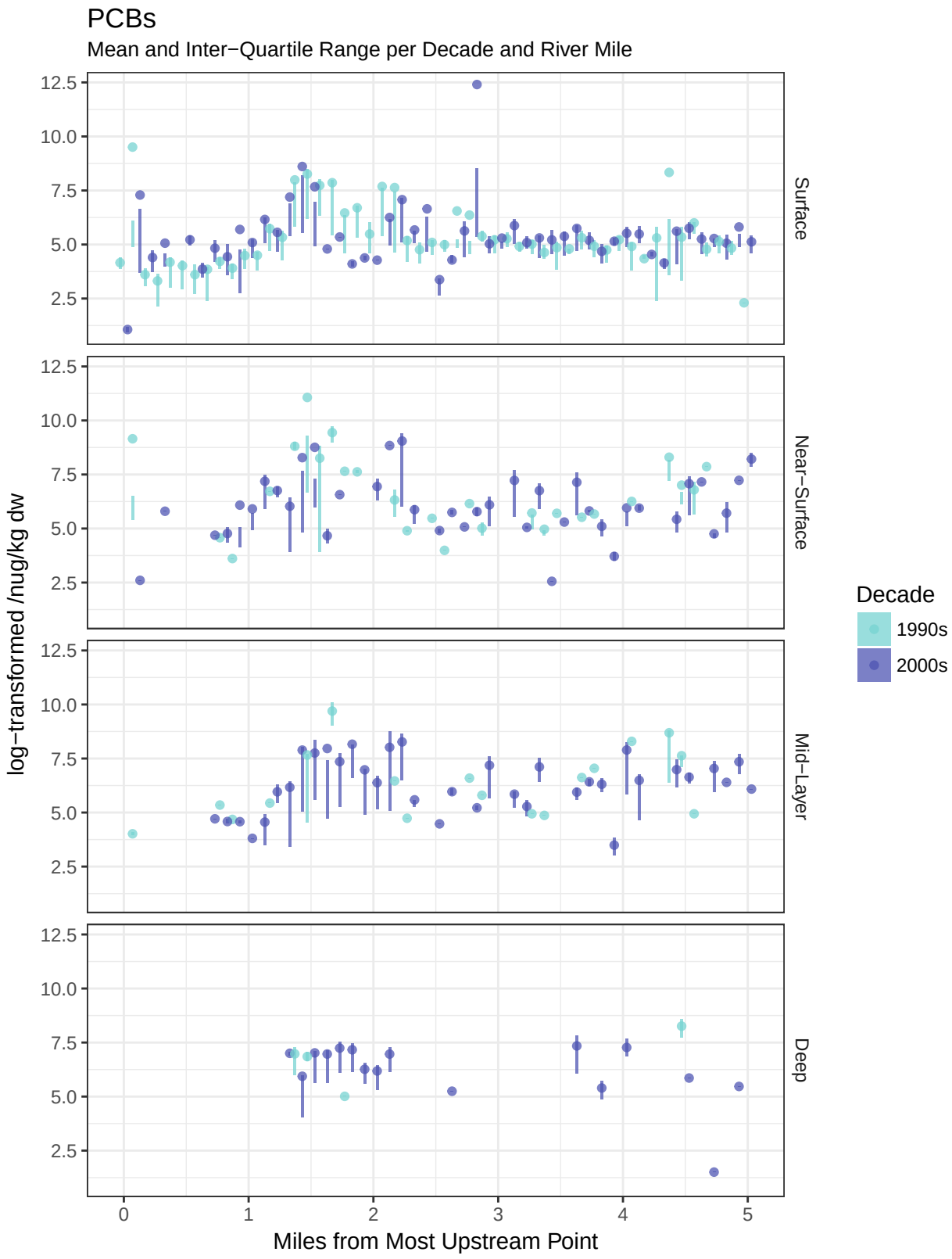
```
#Determine decade
chem_data[Year<2000,Decade:="1990s"]
chem_data[Year>=2000,Decade:="2000s"]

chem_by_decade_rm<-chem_data[!is.na(chem) & Detected=="Yes" &
                             shipping_channel=="N",
                             list(obs_mean=mean(observed_value),
                             obs_sd=sd(observed_value),
                             q1=quantile(observed_value,.25),
                             q3=quantile(observed_value,.75)),
by=c("chem","RM_up","Depth","Decade","Unit")]

# Make a river-mile adjustment (just for the graphics), so that
# both decades can show up in the same plot conveniently even with overlap
chem_by_decade_rm[,adjusted_rivermile:=ifelse(Decade=="1990s",RM_up-.03,RM_up+0.03)]
```

And now generate a series of plots, in which the dots represent the mean, and the bars represent the IQR for the data points.

```
for(chemical in names(chemlist)){
 chem_info<-chemlist[[chemical]]
 #Determine color scheme for plots
 decade_colors<-c("#77D4D2","#4E55B3")
 names(decade_colors) <- c("1990s","2000s")

 # Make plot of highest values
  p<-ggplot() +
  ggtitle(chemical, subtitle="Mean and Inter-Quartile Range per Decade and River Mile")+
    geom_rect(data=chem_by_decade_rm[chem==chemical], # geom_rect for IQR
                  aes(xmin=adjusted_rivermile-.01,
                      xmax=adjusted_rivermile+.01,
                      ymin=log(q1),ymax=log(q3), fill=Decade),alpha=.75) +
         scale_colour_manual(name = "Decade",values = decade_colors, drop=F) +
         scale_fill_manual(name = "Decade",values = decade_colors, drop=F) +
    geom_point(data=chem_by_decade_rm[chem==chemical],# geom_point for mean
                  aes(x=adjusted_rivermile,y=log(obs_mean),
                      color=Decade),alpha=.75) +
    ylab(paste0("log-transformed /n",chem_info[["unit"]]))+
    xlab("Miles from Most Upstream Point") +
    theme_bw()+ facet_wrap(~Depth,ncol=1,strip.position="right")+
    theme(strip.background = element_blank())
  print(p)
}
```


PAHs
Mean and Inter−Quartile Range per Decade and River Mile
log−transformed /nug/kg dw
7.5
5.0
2.5
Surface
Near−Surface
Mid−Layer
Deep
0
1
2
3
4
5
Miles from Most Upstream Point
Decade
1990s
2000s


Arsenic
Mean and Inter−Quartile Range per Decade and River Mile
log−transformed /nmg/kg dw
Surface
Near−Surface
Mid−Layer
Deep
0
2
4
6
0
1
2
3
4
5
Miles from Most Upstream Point
Decade
1990s
2000s


PCBs
Mean and Inter−Quartile Range per Decade and River Mile
log−transformed /nug/kg dw
12.5
10.0
7.5
5.0
2.5
Surface
Near−Surface
Mid−Layer
Deep
0
1
2
3
4
5
Miles from Most Upstream Point
Decade
1990s
2000s

Unfortunately, that is all the exploration and analysis I have time for on this question! However, with these graphs, I can start to get a picture of what might have happened here.

**PAHs**

The contamination from PAHs in the lower Duwamish seems to be characterized by intermixed pollution throughout both the sediment strata and across time, with observed concentrations becoming higher (in general) downstream. Two miles away from the furthest upstream point (or, river-mile 3 as measured from Puget Sound), concentrations seem to increase, with high measurements both deep in the sediment (into the mid-layer and deep sediments). With fewer deep samples in the upstream portion of the river, this apparent trend may be due to sampling bias. High values were observed in the very first samples taken in the mid-1990s, going down to the mid-layer of the sediments, which indicates that the polluting process likely has occurred for years before the samples were taken. Later measurements in the late 2000s point to improving conditions in the surface layer, but the PAH contamination seems to run deep, especially in the lower reaches of the river as it heads towards the sea. Interpreting the graphs by decade, it appears that variation among observations for each tenth-of-a-river-mile strata seems to be relatively consistent across both river mile and decade. The magnitude of the inter-quartile range also points to the relatively noisy observations seen in the plot by river-mile and year.

**Arsenic**

Unlike PAHs, there seems to be very little variation in arsenic measurements, with specific spikes at river-mile locations– a trend that is consistent across decade. At two particular river-mile locations–around 1.2 miles and 3.5 miles from the furthest upstream point, huge spikes of arsenic are observed. Interestingly, the downstream areas do not seem to show a clear "trailing off" from higher values upstream tainting lower portions sequentially, which means that the contamination seems to be relatively isolated. The pollution source for the spike at 3.5 miles must have been present for quite some time– extremely high measurements were found deep in the sediment layers. Another spike just before Puget Sound is present in near-surface, mid-layer, and deep sediments, but was not found on the surface of the river bottom, which could indicate either recent deposition of non-arsenic sediment (after all, the surface measurements are very small in depth, only capturing the very top layer).

**Total PCBs**

Much like arsenic, this contaminant seems to be present in hot-spots along the river's course, rather than widely distributed across the river with noisy measurements like PAHs. Higher values stretch into the sediment's mid-layer at 1.5 miles away from the most upstream measured point (close, but not at the same mile marker as the high arsenic readings, which are slightly upriver). Very high observations in the surface layer around mile 4.5 in the early 1990s, and 2.5 in the mid-2000s could indicate contamination from specific events. Unlike arsenic and PAHs, total PCBs seem to have heteroskedasticity in their measurements– the range of observed values is much wider earlier in the river. This high variance in measurements was present even in the mid-layer and deep sediments– since this analysis takes only river mile, and not bank of the river into account, it's possible that some areas within the same river-mile have widely different contamination levels depending on the closest riverbank.

## Where could I go from here?

These graphs are a good starting point to understand some of the dynamics present in this data set, but they do not provide statistical evidence for certain phenomena, nor do they directly compare measurements across space or time for statistical significance. The "right tests" to undertake would depend on what question we are trying to ask– do we care about upstream-downstream patterns in order to simply map the areas that need interventions, or to attribute downstream cleanup efforts to an upstream source? The driving underlying motivation for this study would undoubtedly shape further analysis.

Some potential leads to follow are as follows:

- To check for whether apparent differences in different river-mile sections were statistically significantly different from others, ANOVA tests could be a good place to start building evidence that one region was substantially worse than the others.
- Explore the X/Y spatial patterns, such as hot spot detection
- Determine the spatial and temporal lags that characterize upstream->downstream auto-correlation

Being relatively unfamiliar with this data type, I would almost certainly be checking with the rest of the team as to the best way to proceed.

### Addressing the bonus: What can you tell about the modification history of the site?

It looks like this data contains bathymetric data coming from one year (2003) based on the column name, regardless of when the sample was taken, so I don't have any information on how the topography of the river bottom may have changed based on this data set. However, this data contains two piece of information that might indicate modification history– both a flag for whether or not the data is within a shipping channel, with the potential for regular dredging, and also a column "DredgeYear", which sometimes (often) is later than the actual sample date– leading me to believe that this column is likely derived from a spatial join to polygon layers describing historical dredge paths. Furthermore, it appears that many of the points with a history of dredging are not within the shipping channel.

Knowing the dredging history doesn't tell me anything about whether sediments or sand were put down to replace the removed material– to look for areas of disturbance or modification, one strategy would be to choose a handful of important chemical analytes within the sediment samples that indicate sediment type, and then look to see whether observations in the same area had shifted substantially in later sediment cores within that region.

As a first pass, however, I can still make a quick map of unique sampled locations, and when (if at all) these samples were dredged:

```
chem_data[DredgeYear%in%c(NA,"","  "),DredgeYear:=" Not Dredged"]
dredge_table<-unique(chem_data[,list(LocationName,X,Y,RM,Dredged,
                                     DredgeYear,shipping_channel)])
dredge_table[,point_index:=seq(1,nrow(dredge_table))]
dredge_history<-MapSuite::PointMap(coords=dredge_table,
                                   id="point_index",xcol="X",ycol="Y",
                                   variable="DredgeYear",
                  map_title="Sample Sites with Historical Dredging",
                  map_subtitle="Lower Duwamish, Washington State",
                  map_colors = wpal("betafish"),
                  legend_position = "bottom",
                  legend_orientation = "horizontal",
                  return_objects = T, include_titles=T,
                  font_size=20, map_transparency=.5)$map
print(dredge_history)
```

# Sample Sites with Historical Dredging

## Lower Duwamish, Washington State