Introduction to exploratory data analysis

Marcus W. Beck[1] Todd D. O'Brien[2]

[1]ORISE, USEPA NHEERL Gulf Ecology Division
Email: beck.marcus@epa.gov

[2]NOAA/NMFS Copepod Project
Email: todd.obrien@noaa.gov

Objectives and agenda

- Objectives
  - What are some tools for pre-processing/organizing the SWMP data?
  - What is the purpose of exploratory data analysis (EDA)?
  - What are some common techniques and tools for EDA?

Objectives and agenda

- Objectives
  - What are some tools for pre-processing/organizing the SWMP data?
  - What is the purpose of exploratory data analysis (EDA)?
  - What are some common techniques and tools for EDA?
- Agenda
  - Organizing tools in SWMPr
  - Purpose and overview of EDA
  - Generic EDA tools in R, tools in SWMPr

Interactive portion

You can follow along in this module:

- dataset2
- script2

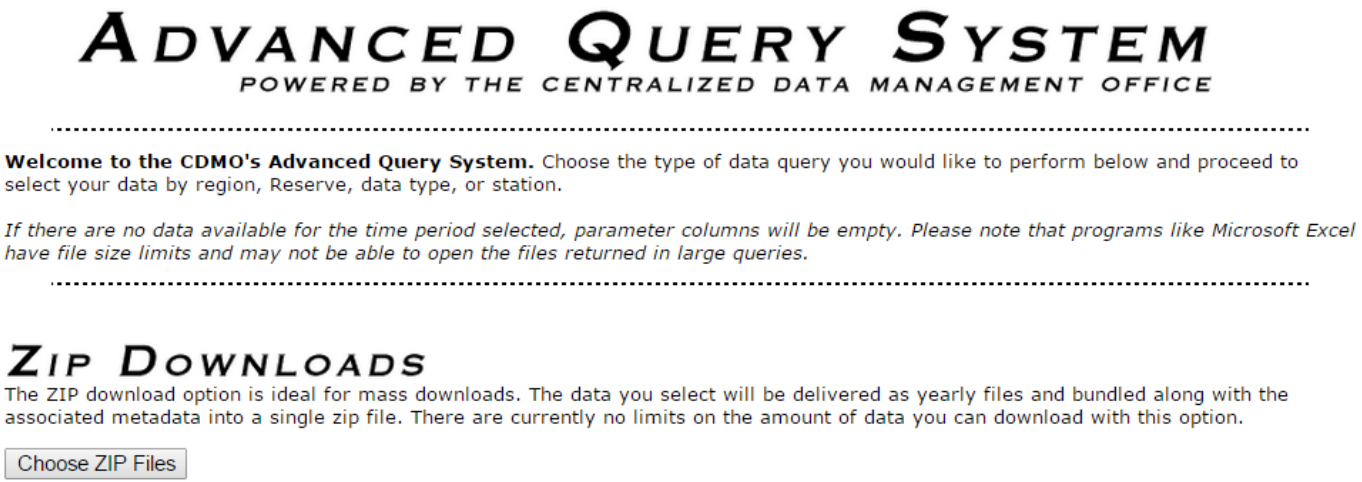
Interactive! Interrupt me!

Import SWMP data and organize

We learned how to import SWMP data in the previous session

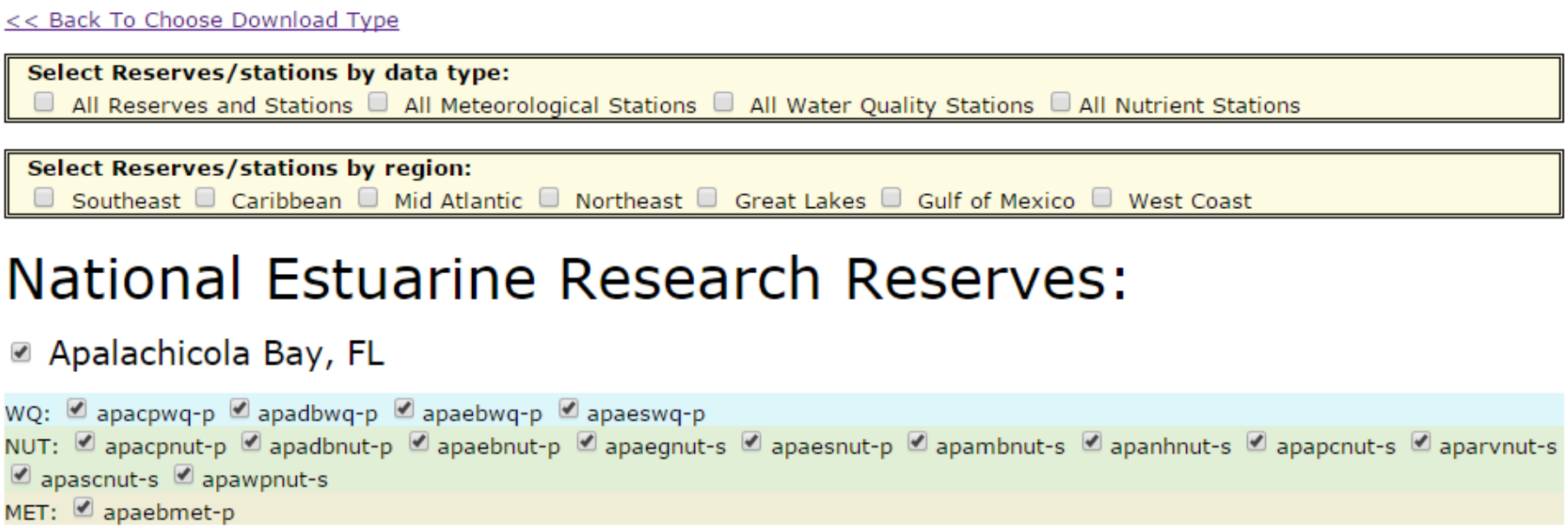
To review, the easiest approach is to download the data outside of R, then import using the 'import˙local' function

Be sure that you use only the zip downloads feature from CDMO - the 'import˙local' functions works best with these data



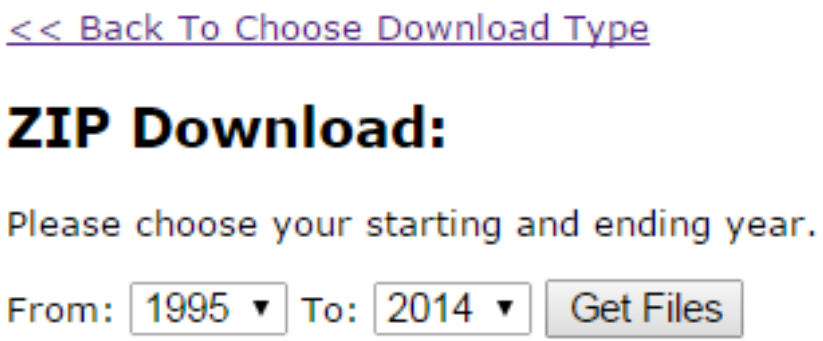
Import SWMP data and organize

It's best to download all the data possible for a reserve to avoid repeated requests to the server and to centralize the location from which the data are imported into R



Import SWMP data and organize

It's best to download all the data possible for a reserve to avoid repeated requests to the server and to centralize the location from which the data are imported into R



Here we've made a request for all stations at Apalachicola Bay (water quality, nutrients, weather) and all available years (1995–2014)

This request will take several minutes to be delivered to your email - an abbreviated version of these data are provided with the workshop materials for this training module

Import SWMP data and organize

Let's import some data for Apalachicola Bay

```
# reload the SWMPr package just in case
library(SWMPr)

# import data
# change this path for the flash drive
path <- 'C:/data/dataset2'
wq_dat <- import_local(path, 'apacpwq')
nut_dat <- import_local(path, 'apacpnut')
met_dat <- import_local(path, 'apaebmet')
```

We've just imported data from 2011–2014 for three stations (apacpwq, apacpnut, apaebmet) and saved them in our workspace as three separate objects (wq˙dat, nut˙dat, met˙dat)

Import SWMP data and organize

But don't take my word for it, take a look at the data!

```
# what are the dimenions of the water quality data?
dim(wq_dat)

## [1] 132035     25

# what are the dimenions of the nutrient data?
dim(nut_dat)

## [1] 48 13

# what are the dimenions of the weather data?
dim(met_dat)

## [1] 133548     23
```

Import SWMP data and organize

View the first six rows

```
# View the first six rows of the wq data
head(wq_dat)
```

```
##         datetimestamp temp f_temp spcond f_spcond sal f_sal do_pct f_do_pct
## 1 2011-01-01 00:00:00   11   <0>      44     <0>   28  <0>      68     <0>
## 2 2011-01-01 00:15:00   11   <0>      44     <0>   28  <0>      68     <0>
## 3 2011-01-01 00:30:00   11   <0>      44     <0>   28  <0>      68     <0>
## 4 2011-01-01 00:45:00   11   <0>      44     <0>   28  <0>      68     <0>
## 5 2011-01-01 01:00:00   11   <0>      44     <0>   29  <0>      68     <0>
## 6 2011-01-01 01:15:00   11   <0>      44     <0>   29  <0>      67     <0>
##   do_mgl f_do_mgl depth f_depth cdepth f_cdepth level f_level clevel f_clevel
## 1      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
## 2      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
## 3      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
## 4      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
## 5      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
## 6      6     <0>      2    <0>       2     <3>     NA   <-1>      NA        NA
##   ph f_ph turb f_turb chlfluor f_chlfluor
## 1  8 <0>     3   <0>        NA      <-1>
## 2  8 <0>     3   <0>        NA      <-1>
## 3  8 <0>     2   <0>        NA      <-1>
## 4  8 <0>     1   <0>        NA      <-1>
## 5  8 <0>     2   <0>        NA      <-1>
## 6  8 <0>     1   <0>        NA      <-1>
```

Import SWMP data and organize

View the last six rows

```
# View the last six rows of the wq data
tail(wq_dat)
```

```
##              datetimestamp temp f_temp spcond f_spcond sal f_sal do_pct
## 132030 2014-10-07 07:45:00   24   <0>      41     <0>   26  <0>      90
## 132031 2014-10-07 08:00:00   24   <0>      41     <0>   26  <0>      91
## 132032 2014-10-07 08:15:00   23   <0>      39     <0>   25  <0>      95
## 132033 2014-10-07 08:30:00   23   <0>      39     <0>   25  <0>      95
## 132034 2014-10-07 08:45:00   24   <0>      38     <0>   24  <0>      95
## 132035 2014-10-07 09:00:00   24   <0>      38     <0>   24  <0>      96
##        f_do_pct do_mgl f_do_mgl depth f_depth cdepth f_cdepth level f_level
## 132030     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
## 132031     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
## 132032     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
## 132033     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
## 132034     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
## 132035     <0>       7     <0>      2    <0>       2     <3>     NA   <-1>
##        clevel f_clevel ph f_ph turb f_turb chlfluor f_chlfluor
## 132030     NA       NA  8 <0>    10   <0>       NA       <-1>
## 132031     NA       NA  8 <0>     8   <0>       NA       <-1>
## 132032     NA       NA  8 <0>     7   <0>       NA       <-1>
## 132033     NA       NA  8 <0>     7   <0>       NA       <-1>
## 132034     NA       NA  8 <0>     7   <0>       NA       <-1>
## 132035     NA       NA  8 <0>     5   <0>       NA       <-1>
```

Import SWMP data and organize

What class is the data?

```
# class of the data
class(wq_dat)

## [1] "swmpr"      "data.frame"
```

This tells us that the imported data are two different classes - 'swmpr' and 'data.frame'

The class of an object is important because it defines the types of methods (i.e., functions) that apply

For example, 'head' and 'tail' functions apply to a 'data.frame'

Import SWMP data and organize

The swmpr object class was developed to make your life easier working with SWMP data, i.e., functions in the SWMPr package organize and analyze raw SWMP data

The online documentation describes the functions that work with the swmpr object class, also...

```
# what functions/methods work with swmpr objects?
methods(class = 'swmpr')
```

```
##  [1] aggregate.swmpr comb.swmpr      decomp.swmpr    hist.swmpr
##  [5] lines.swmpr     na.approx.swmpr plot.swmpr      qaqc.swmpr
##  [9] qaqcchk.swmpr   setstep.swmpr   smoother.swmpr  subset.swmpr
```

Documentation of each function can be viewed as follows (although currently not complete):

```
# see help for a swmpr function
?aggregate.swmpr

# or...
help('aggregate.swmpr')
```

Import SWMP data and organize

A useful feature of R is that a defined class will have both ***data*** and ***attributes***

For the swmpr object class, the ***data*** are the raw swmpr data as a data.frame

The ***attributes*** are a list of metadata for the imported data

```
# what attributes are available for a swmpr object
names(attributes(wq_dat))

## [1] "names"       "row.names"   "class"       "station"     "parameters"
## [6] "qaqc_cols"   "date_rng"    "timezone"    "stamp_class"

# view the parameters
attr(wq_dat, 'parameters')

##  [1] "temp"     "spcond"   "sal"      "do_pct"   "do_mgl"   "depth"
##  [7] "cdepth"   "level"    "clevel"   "ph"       "turb"     "chlfluor"
```

Import SWMP data and organize

You can also view all the attributes as follows:

```
# view all attributes
attributes(wq_dat)
```

This is not recommended since they are quite long, e.g., an attribute of the 'data.frame' class is the row names (132035 rows for'wq_dat')

Individual attributes are useful for getting a feel for the dataset - what is the date range? what parameters are included? are QAQC columns present?

However, the intended use of attributes is behind the scenes with swmpr functions - they will be used to process the data and updated automatically

Import SWMP data and organize

Now that we have a feel for the data, what needs to be done before we can start analyzing the information?

Last module:

- How do we handle QAQC data or 'bad' observations?
- How do we deal with data we don't want?
- How do we combine data for comparison?
- How do we handle issues inherent with time series?

Several of these problems are context-dependent - driven by the question or analysis

Others are common to any analysis...

Import SWMP data and organize

Perhaps the first organizational tool you want to use is 'qaqc.swmpr'

This function does two things:

- Remove observations with a specified QAQC flag value
- Remove extraneous QAQC columns

-5 Outside high sensor range
-4 Outside low sensor range
-3 Data rejected due to QAQC
-2 Missing data
-1 Optional SWMP supported parameter
0 Passed initial QAQC checks
1 Suspect data
2 Open - reserved for later flag
3 Calculated data: non-vented depth/level sensor correction for changes in barometric pressure
4 Historical data: Pre-auto QAQC
5 Corrected data

Import SWMP data and organize

You will have to decide which values to keep - be conservative and only keep those that passed QAQC (best option?) or keep all the data (worst option?)

To help you decide, it may be useful to get an idea of the distribution of QAQC flags in the data

```
# use qaqcchk to view distributin of qaqc flags
myqaqc <- qaqcchk(wq_dat)

# view first six rows
head(myqaqc)
```

```
##        piece f_cdepth f_chlfluor f_depth f_do_mgl f_do_pct f_level f_ph f_sal
## 1                288          NA       NA        NA        NA       NA   NA    NA
## 2       <-1>      NA      132035       NA        NA        NA   132035   NA    NA
## 3 <-1> [GCU]       9          NA       NA        NA        NA       NA   NA    NA
## 4       <-2>      NA          NA       90        90        90       NA   90    90
## 5 <-2> (CSM)      NA          NA       51        51        51       NA   51    51
## 6 <-2> [GCM]    1353          NA       NA        NA        NA       NA   NA    NA
##   f_spcond f_temp f_turb
```


NERRS / SWMP
Data Analysis Workshop: Time Series
November 17, 2014

Questions??