Purpose: working framework to the DS health and conditions dataset EDA analysis

1. Begin EDA
   1. Explore dataset
      1. Metadata is sparse: document and expand metadata
         1. Assess missing data
            1. Total missing

Chart, line chart

AI-generated content may be incorrect.

Most variables are missing

Graphical user interface, application, table, Excel

AI-generated content may be incorrect.

Full dataset data

hsi, cf, and cf\_no\_gonads potential response variables

No real predictor variables span entire dataset

* + - * 1. Patterns?

Would the NA distribution change if focusing on only the summer/fall?

Might not be worth pursuing given there is likely a lag effect on predictors to health metrics

* + - 1. What can be potential response vs predictor variables?
         1. Response variables:

cf, and cf\_no\_gonads, hsi

Chart, line chart

AI-generated content may be incorrect.

Cf more normal than hsi (log/sqrt transformations didn’t help much)

May be more normal if you removed values from adults though

cf\_no\_gonad, remove confounding factor of maturation?

Inconsistent in subadult fish, data quality concerns (Bruce)

* + - * 1. Predictor variables

Start with water quality variables from the sampling trawls themselves

Joined in water quality data from each survey (SKT, EDSM, STN, FMWT)

By fish id as first layer, station and sampling date as second layer

Some issues with EDSM data; the FL recorded per fish id in the health dataset != those by EDSM

Incorporate dayflow data

Incorporate CDEC data

Per EDSM subregion, mean daily water quality data:

Chart, radar chart

AI-generated content may be incorrect.

All CDEC station within these subregions

Find stations with the following data

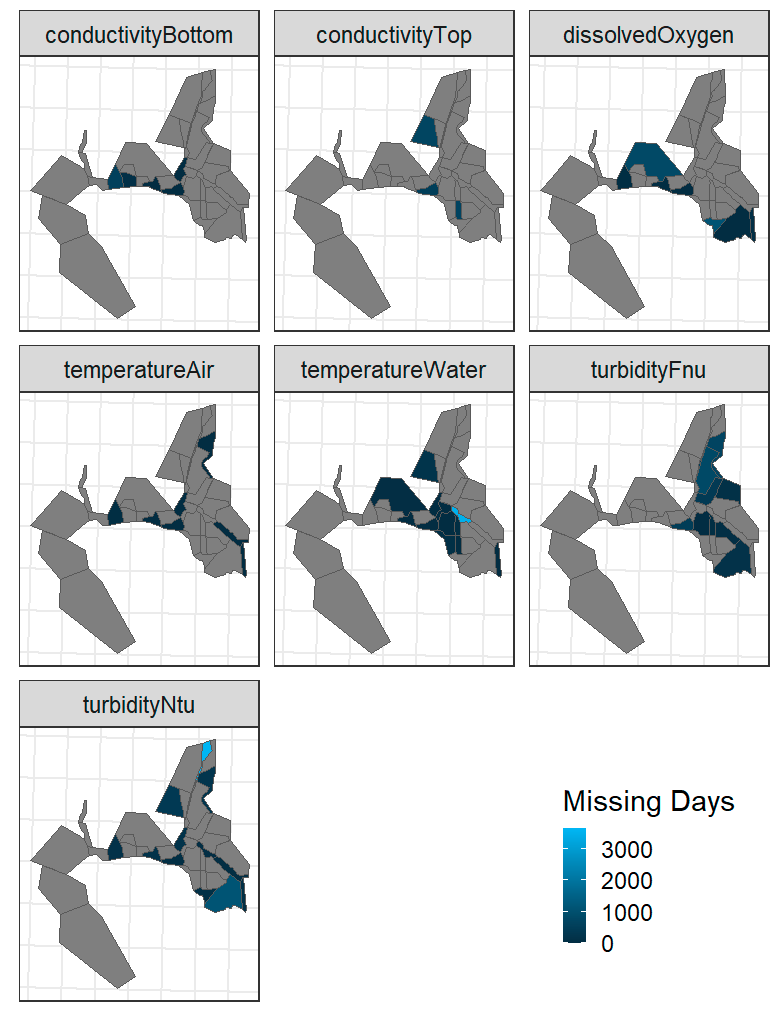
Water temperature

Air temperature

Conductivity

Turbidity

DO



Ask for smaller span of data, yearly

Limited data

Chart, radar chart

AI-generated content may be incorrect.

Can aggregate to strata level instead of subregion

Can aggregate to higher than daily level data

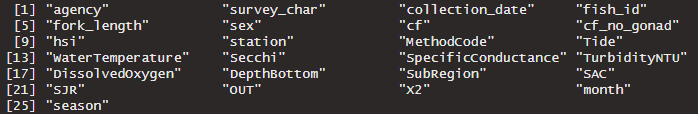
May need to incorporate in USGS, NOAA, discrete data? Worth?

Min distance to GGB could be a proxy for salinity?

Chart, surface chart

AI-generated content may be incorrect.

* + - * 1. Fish catch per region? Does using subregions even make sense? Do we need all of these subregions?
        2. Current variables



* + - 1. Own definition of a usable variable: theoretical motivation to including it
         1. Response:

hsi, liver/total weight, influenced by maturation

Big difference between “sex”

“sex” data sparse for non-adult fish

cf, weight/length, could be influenced by gonad development

* + - * 1. Predictors:

Water temperature: affects growth rate, survival

Maximum threshold or a measure of time spent above optimal range

Air temperature: correlation to water temperature if can help data imputation

Secchi/turbidity: a proxy for survival (feeding, avoidance)

Conductance: habitat activation

Dissolved oxygen: survival

SAC, SJR, OUT, X2: outflow metrics, habitat activation, flow-related effects

Month/season: proxy for maturation

Year: population trend

Survey: sampling method

* + 1. Maximizing usable data:
       1. Test some thresholds, only accept variables with enough data?
          1. How do we make this threshold less ambiguous? Could identify the potential predictors and define a threshold that would include them all?

This would include predictors that we think might not be useful but can be utilized in the model and importance determined

* + - * 1. Only the response variables are usable from the dataset

All other predictors have been integrated in

Others?

* + - 1. Maximize correlation and minimize Nas
    1. What does the response variable represent?
       1. Greater fish health equates to what? Greater growth? Greater survival?
       2. Which variable is the best candidate?
          1. Minimize NAs, outliers
          2. Normalizable distribution
    2. What can be used as a predictor variable?
       1. Are there predictors that represent “good conditions” that “benefits” DS?
          1. Need to think through how to aggregate (temporal and spatially) or engineer certain predictors (e.g., % of days beyond range)
       2. Can I supplement predictors with other data, i.e., data from other surveys or continuous sensor data, to fill in NAs?
          1. Done this first part, although can add more
          2. Does imputation make sense? Running averages?
       3. Relationship to the response: plot them
          1. Chart, line chart

             AI-generated content may be incorrect.

Curious to how there is a negative correlation in X2 after holding seasonality constant

* + - * 1. No relationships to cf; some relationships to hsi but only because skewed by maturation of fish

Diagram

AI-generated content may be incorrect.

Calendar

AI-generated content may be incorrect.

HSI only: Adult fish in 2017 causing the correlation

Need to take out these adult fish, but likely will see that there is no relationship with instantaneous data

* + - * 1. Need for temporal aggregations and lags

What’s appropriate?

Spatial aggregation likely will be important as well, but does data exist?

* 1. Preliminary models
     1. Select potential models based on the chosen response variable and data characteristics
     2. Do predictors need to be transformed?
     3. Start with a correlation analysis
        1. 0.65 threshold for collinearity
           1. VIF > 5/10 can work as well
        2. Scatterplot/boxplots to visualize relationship
        3. What are the theoretical basis for why important predictors are correlated?
           1. Composition feature made?

1. Modeling