**Next Steps 11/21:**

**Three questions：**

1. What is infrastructure construction’s effect on biodiversity (species abundance)? In other words, count of species over phases at diff distances
2. Did the species composition change at diff sites over phases? Especially looking at closer sites
3. Did distribution of size (mass & body length) changes over phases at diff distances?

**To do list:**

1. **Dec. 5** Try to come up with meaningful/useful analysis

* Select at least one of the questions listed above and perform data analysis (using at least 2 different methodologies)
  + Possible analysis:
    - Random Forest to predict Phase
      * Binary class (abandon vs. the rest)
      * Multiclass
      * One vs all
    - PERMANOVA
    - Occupancy Modeling
    - Linear Reg
    - T test and or chi-square test
    - PCA / canonical correspondence analysis (CCA) or Non metric multidimensional scaling (NMDS)
* Create new Viz for analysis
  + Viz (heat geomap)
    - Show abundency
    - Can set drop down menu for selecting different phases/ use different shape to represent different phases
  + ROC
* Select 3~5 of the representative visualization from EDA and make adjustment (one person)
  + Unify units, attributes, size, Viz title, etc.

**Meeting with Xi 11/21**

Random Forest to predict Phase:

Main idea: if the prediction accuracy is high, it means the features we used for prediction actually has a strong correlation with construction phase

* Using attributes that Jessie mentioned to run Random Forest (scikit learn)
* Try to tune parameter
* Perform feature ranking
* Remember to normalize the data before training

If multiple different analysis can draw to the same conclusion, then it is safe to make the conclusion that the construction does have an impact on amphibian

Xi’s note:

1. Can perform stratified cross validation during training
2. If you are interested and have time, you can play the data with LGB. It may not provide better results, but it runs faster and it’s always good to learn new things.
3. Some knowledge about petroleum drilling:
   1. EC and DR phase usually have a larger instant impact on the environment
      1. EC: loud noise when cutting down trees
      2. DR: when drilling starts, the large equipment never stops, it will work day and night, may also produce loud noise
   2. Such construction may affect water quality

**Next Steps 11/07:**

1. Clean data set for everyone to use and keep track of the changes (Masha to send around, add 2016 data)
2. Three questions
   1. What is infrastructure construction’s effect on biodiversity (species abundance)? In other words, count of species over phases at diff distances
      1. Viz (heat map)
   2. Did the species composition change at diff sites over phases? Especially looking at closer sites
   3. Did distribution of size (mass & body length) changes over phases at diff distances?

Analyses:

Random Forest to predict Phase

PERMANOVA

Occupancy Modeling

Linear Reg

T test and or chi sq test

**3 rd meeting 11/07**

Meeting with Jessica:

Resampling technique

PERMANOVA for composition

Visualize the results for each of the three questions

Writing up the methods (R, python and packages)

Enough detail for someone without technical background

PCA for count (aka abundance) it doesn’t not give the best result

A diff coordinate analysis may be better - canonical correspondence analysis (CCA) or Non metric multidimensional scaling (NMDS) (used even more bc it lets you use the number of dimensions)

End product - produce smth and she will pick out the most relevant

This month - between now and thanksgiving she is v busy

**Masha:**

1. how did construction by diversity by distance

dist/species/plot/phase

2. did family composition change by different plots

3. dist of size(mass) of the species at different distance

phase and counts by distance

statistical: t test, resampling

predictive: features that related to

the end product

what outputs are expected

specially equally discover/ hidden

**EDA 10/25**

Questions for Jesse:

Average size by species/families

Lat/Long issue for a lot of the data - are there diff records we can look from?

**Questions:**

1. How were the animals (distance from the drilling site) affected by the noise? Did they move away?
   1. No of observations by family by phase vs. No of workers/machines/flights
   2. Bin into size groups vs No of workers/machines/flights
2. Hypothesis testing:
   1. Diff distance / diff time / diff yearly — significant difference between SVL/Mass for all of the observations and grouped by family/species/genus
   2. Compare our data with avg data for species/families from Jesse