

## Data Cleaning and Processing

- Convert coordinates from Universal Transverse Mercator (UTM) to Geographic (latitude, longitude) coordinate system
  - Transformation code file: coordinates conversion.R
  - Data file: CleanData\_v3.csv
- Create the abundance dataset for species composition analysis
  - Transformation code in the composition analysis R file
  - Data file: CompositionData\_v1.csv
- Create a visualization dataset with scattered geo locations for observations
  - Transformation code file: Scattered\_Geolocation\_Transform.R
  - Data file: vis\_data\_v1.csv

## Analytical Methodologies

### Non-metric Multidimensional Scaling (NMDS)

#### Purpose

Similar to PCA, NMDS represents the original position of data in multidimensional space as accurately as possible using a reduced number of dimensions that can be easily plotted and visualized. This is a rank-based approach.

#### Assumptions

- Transforming the dataset is desired.
- Can't handle missing values or negative data.
- This method only provides observations. To test whether the samples are statistically different based on grouping, use the ANOSIM test.

#### Inputs

All the predictors.

#### Results and Insights

The five different classes were differentiated very well, which we thought might be an overfitting situation. So we removed all the human factors and rerun the model. But unfortunately, R produced some errors which we couldn't resolve right now. NMDS is usually used as a previous step for ANOSIM test. We proceeded and found that ANOSIM might not be the best test in this case. The detailed analysis is written below.

### Analysis of Similarities (ANOSIM)

#### Purpose

ANOSIM tests statistically whether there is a significant difference between groups of sampling units. If two groups of sampling units are really different in their species composition, then compositional dissimilarities between the groups ought to be greater than those within the groups (ANOSIM R Documentation).

#### Assumptions

- ANOSIM does not assume equal group variances (unlike Permanova)
- **Only allows one variable model**
- The statistical significance of observed R is assessed by permuting the grouping vector to obtain the empirical distribution of R under null-model
- Multiple ways to calculate dissimilarity. In this case, we use Bray-Curtis dissimilarity.

## Inputs

- Abundance data (one with all species and the other with only species which have 10+ observations)
- Group factors (for separate anosim test): phase, distance

## Results

For each dataset (one includes all species and the other one with species which have counts greater than 10), we conducted two separate tests with grouping factors as phase and distance, respectively. Every test generates an R value between 0.1 - 0.3, indicating the dissimilarity among phase groups and among distance groups are small.

(Note: An R value close to "1.0" suggests dissimilarity between groups while an R value close to "0" suggests an even distribution of high and low ranks within and between groups. )

## Insights

The ANOSIM tests generate less robust results. One reason is ANOSIM is incapable of incorporating the interaction effect between phase and distance. We may need to seek the help of other tests. We believe the ANOSIM test is not particularly useful in this case.

## Canonical Correspondence Analysis (CCA)

### Purpose

For random variable X and Y, CCA will find linear combinations of X and Y which have maximum correlation with each other.

### Assumptions

- A group of X and A group of Y.
- Samples need to be random and independent.

### Inputs

Y: "Genus", "Distance", "Location.Type"

X: "Height", "Behavior", "Mass", "Body.Length", "OPE", "Rain", "S\_fuel", "S\_hour", "S\_hr.pp", "S\_peeps", "flights", "Temp", "Luna", "Elev", "Long", "Lat", "Northings", "Eastings"

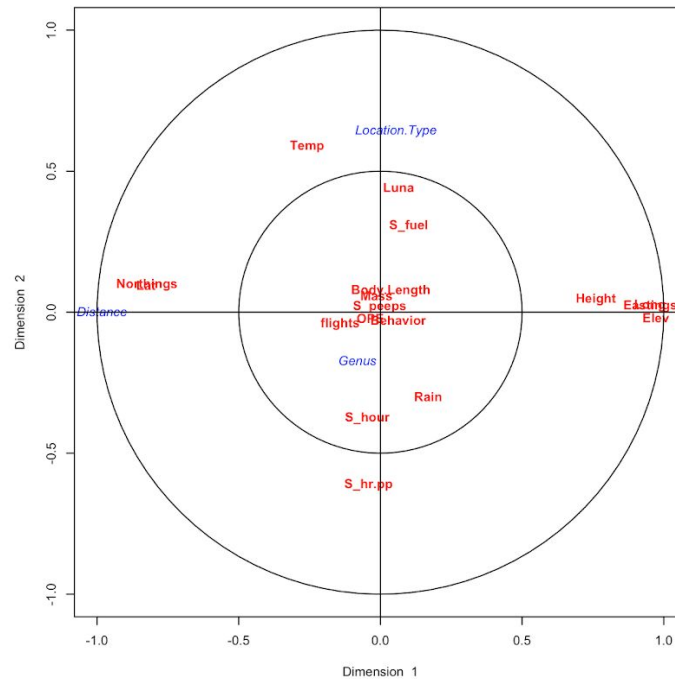
### Results

1. #correlations between two X matrices and each Y  
[1] 0.9835876 0.6667611 0.3830609
2. Regression coefficient between x and y

```
> cc1[3:4] # like regression coefficient
$xccoef
      [,1]      [,2]      [,3]
Height -1.320795e-04  5.624610e-04  3.088241e-03
Behavior -7.354169e-04  1.294452e-02 -1.094353e-01
Mass 1.080371e-02  1.798075e-02  1.078289e-01
Body.Length -5.578844e-04  4.401897e-03 -1.727823e-02
OPE 4.741836e-02  4.006594e-01  6.291113e-02
Rain -1.868674e-03 -1.973280e-02  1.739397e-02
S_fuel 1.261633e-04  1.893867e-03  1.813064e-03
S_hour -4.566287e-04 -4.232869e-03 -9.157679e-05
S_hr.pp 2.520704e-02  8.693461e-02  4.466403e-02
S_peeps 1.014718e-03  1.630728e-02 -8.145272e-03
flights -3.302267e-03 -1.301972e-02 -1.344930e-02
Temp -1.101554e-02  1.073078e-01  6.175827e-02
Luna -1.068334e-01 -5.352022e-01 -5.251642e-02
Elev 6.430065e-03  3.350281e-03 -1.529859e-02
Long -6.415419e+05  8.231583e+06  6.175252e+06
Lat -6.221978e+06 -1.628704e+06 -9.934475e+06
Northings 5.627510e+01  1.412155e+01  8.932679e+01
Eastings 5.470134e+00 -7.597440e+01 -5.761606e+01

$ycoef
      [,1]      [,2]      [,3]
Genus -0.0025823722 -0.0324252335  0.2795443847
Distance -0.0031388116  0.0001977692 -0.0002151156
Location.Type -0.0002991568  0.0587294138  0.0156254242
```

3. Variable representation of the canonical variates, which are the new variables (variate) formed by making a linear combination of two or more variates (variables) from a data set.
4. This displays the maximized correlations between transformed variables of X and Y



### Insights

We selected variable Genus, Distance, and Location.Type as the group of response variables. We could get a basic idea between each  $X_i$  to Y based on linear regression coefficient. The plot above shows that if the variables are further away from the origin meaning that these variables have higher correlations. We are interested in the proposition of the variables. The human factors S\_hour and S\_hr.pp are pointing downwards. The Temperature, Location.Type, Luna and S\_fuel are pointing upwards. The position factors such as latitude, height and elev are squeezed on the right. We could conclude that the propositions and human factors are the dominant factors that influence the Genus, Location.Type and Distance.

## Partial Dependence Plot (PDP)

### Purpose

The partial dependence plot (PDP) shows the marginal effect one or two features have on the predicted outcome of a machine learning model. PDPs show how a feature affects predictions. PDP can show the relationship between the target and the selected features via 1D or 2D plots.

### Assumptions

1. PDP is plotted based on a fitted model.
2. The number of variables shown in PDP could only be 1 or 2.

### Interpretation:

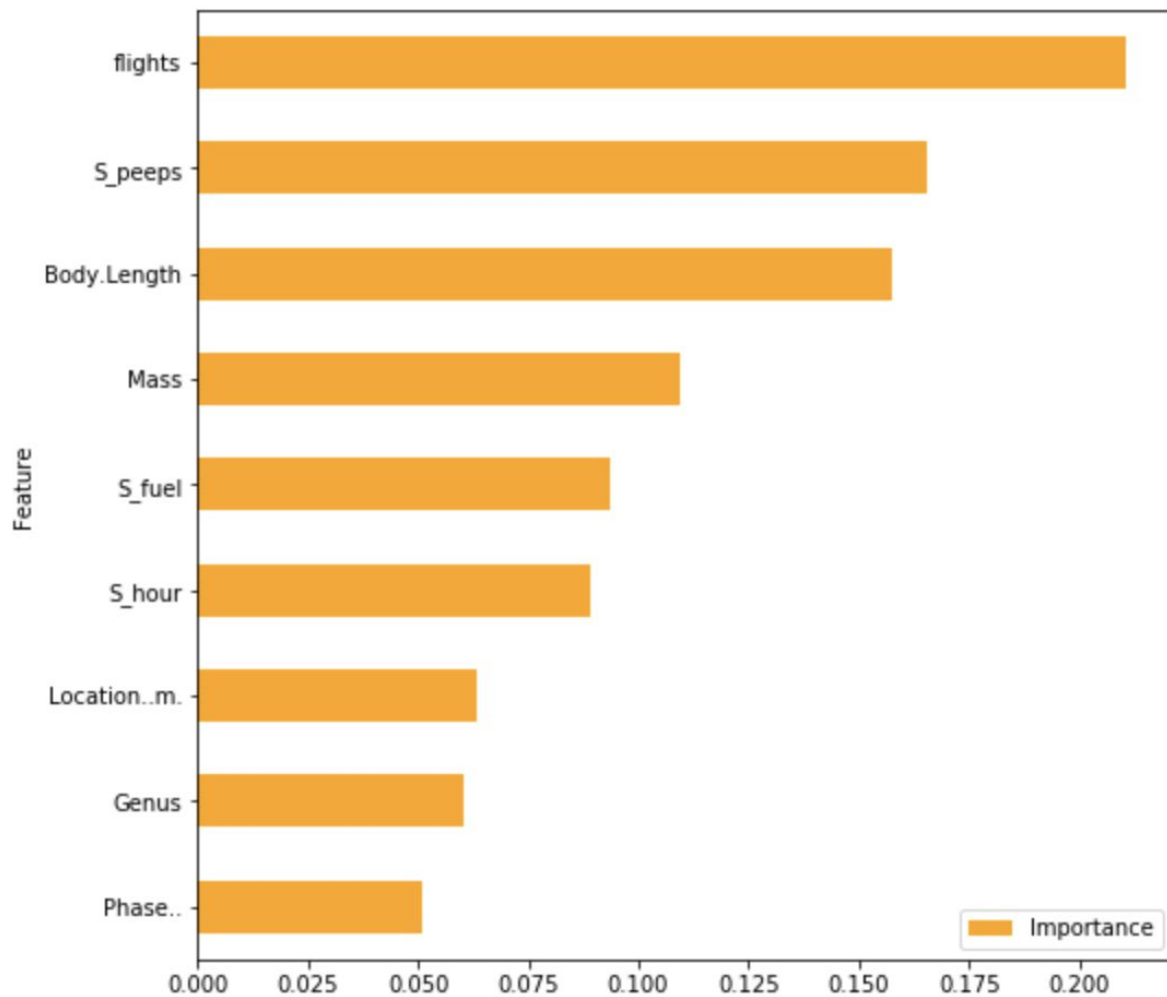
1. The Y-axis represents the change in prediction from what it would be predicted at the baseline or leftmost value.
2. Blue area denotes the confidence interval.

### Inputs

Y: 'Distance'

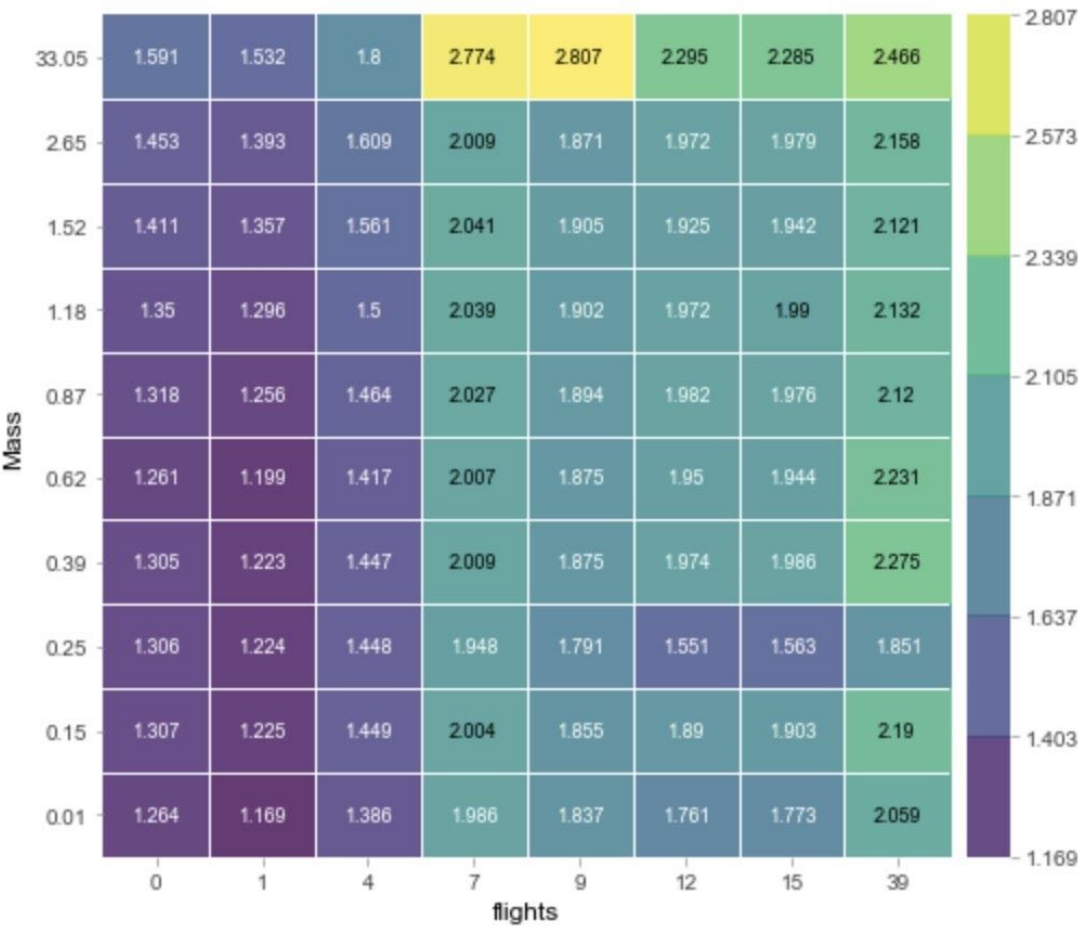
X: 'Phase..', 'Distance', 'Genus', 'Location..m.', 'Mass', 'Body.Length', 'S\_fuel', 'S\_hour', 'S\_peeps', 'flights'

### Results



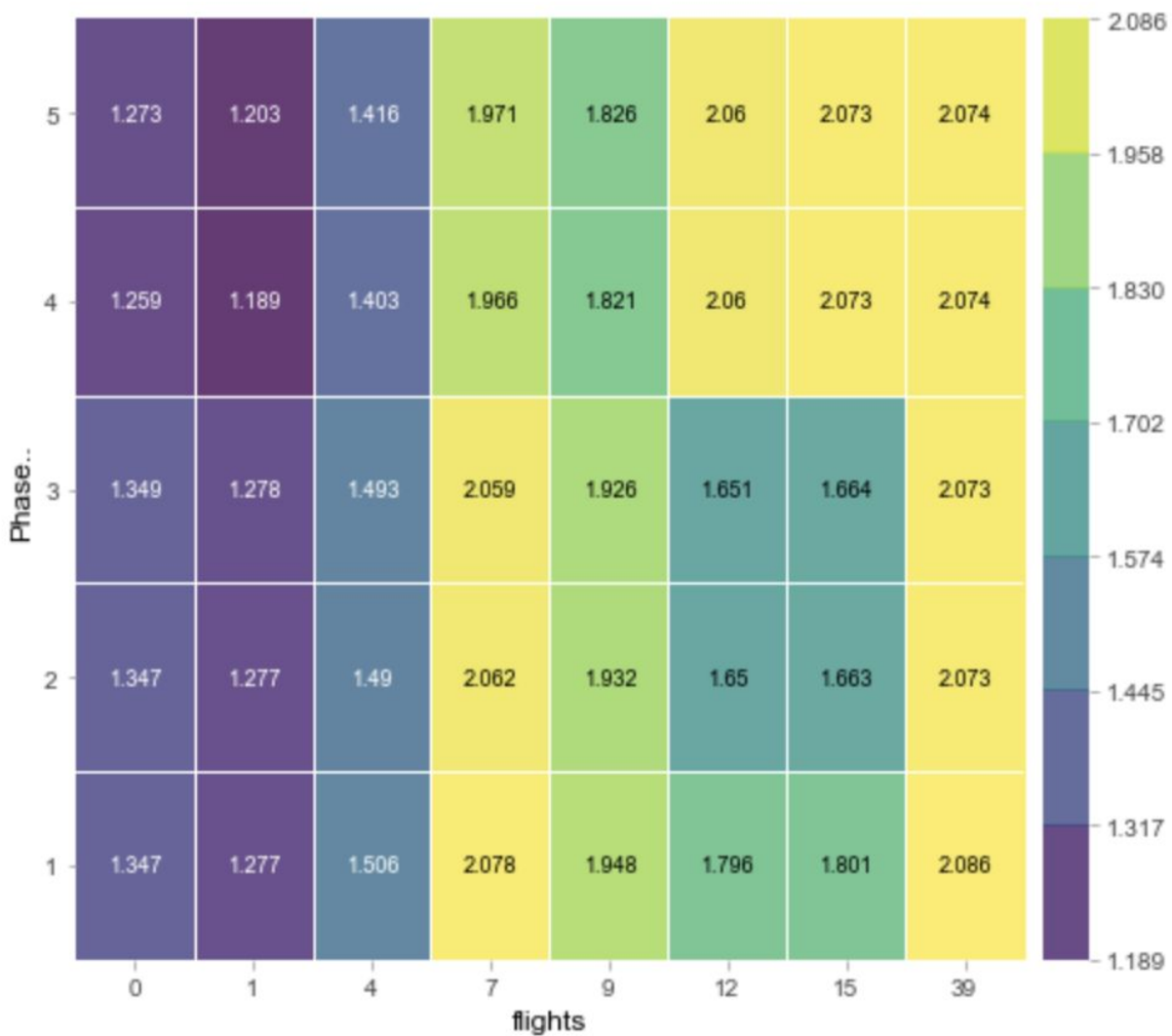
PDP interact for "flights" and "Mass"

Number of unique grid points: (flights: 8, Mass: 10)



## PDP interact for "flights" and "Phase.."

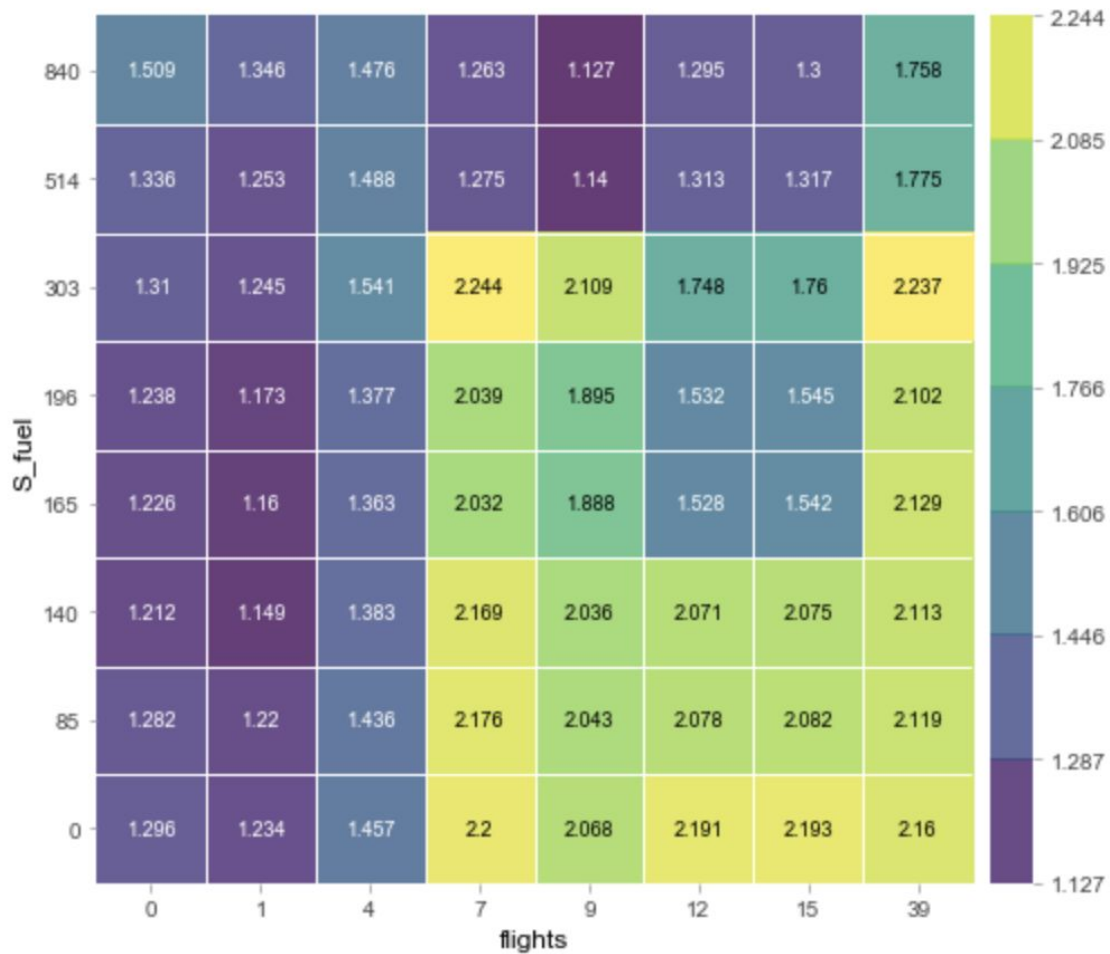
Number of unique grid points: (flights: 8, Phase..: 5)





### PDP interact for "flights" and "S\_fuel"

Number of unique grid points: (flights: 8, S\_fuel: 8)



### Insights

1. Flights, S\_peeps and Body Length are the most important factors that have influence on Distance.
2. Generally, the interesting phenomenon we found from the model is that S\_fuel is negatively correlated to distance.
3. Phase is not an important predictor while the variations of human factors in each phase such as flight, S\_peeps, S\_fuel matter.

## Sharpley Value

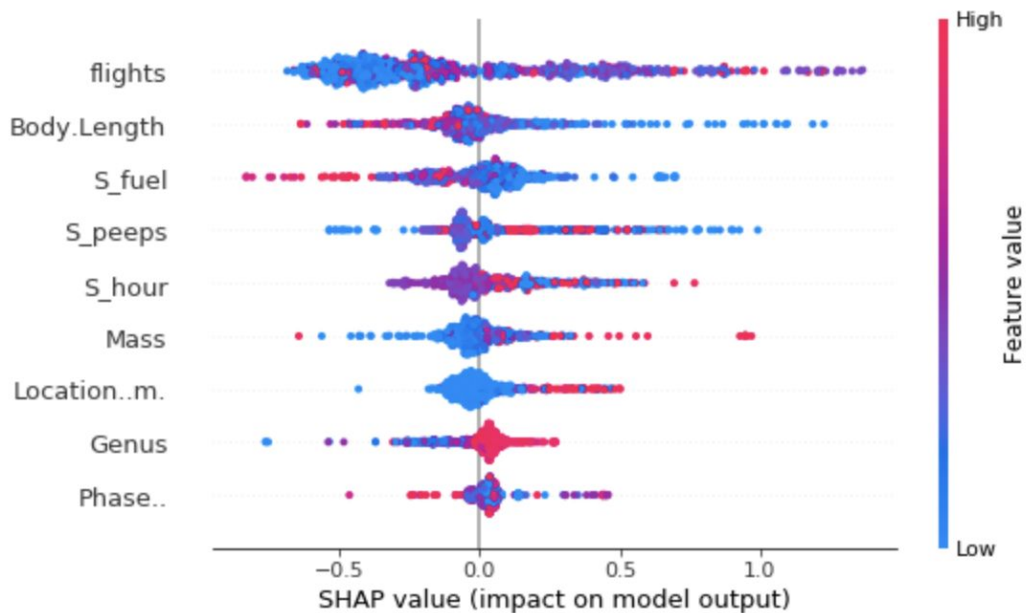
### Purpose

SHAP which stands for Shapley Additive explanation, helps to break down a prediction to show the impact of each feature. It is based on Shapley values, a technique used in game theory to determine how much each player in a collaborative game has contributed to its success. Normally, getting the trade-off between accuracy and interpretability just right can be a difficult balancing act but SHAP values can deliver both.

### Model

The same as PDP.

### Result



The above SHAP summary plots tell which features are most important, and also their range of effects over the dataset.

For every dot:

Vertical location shows what feature it is depicting

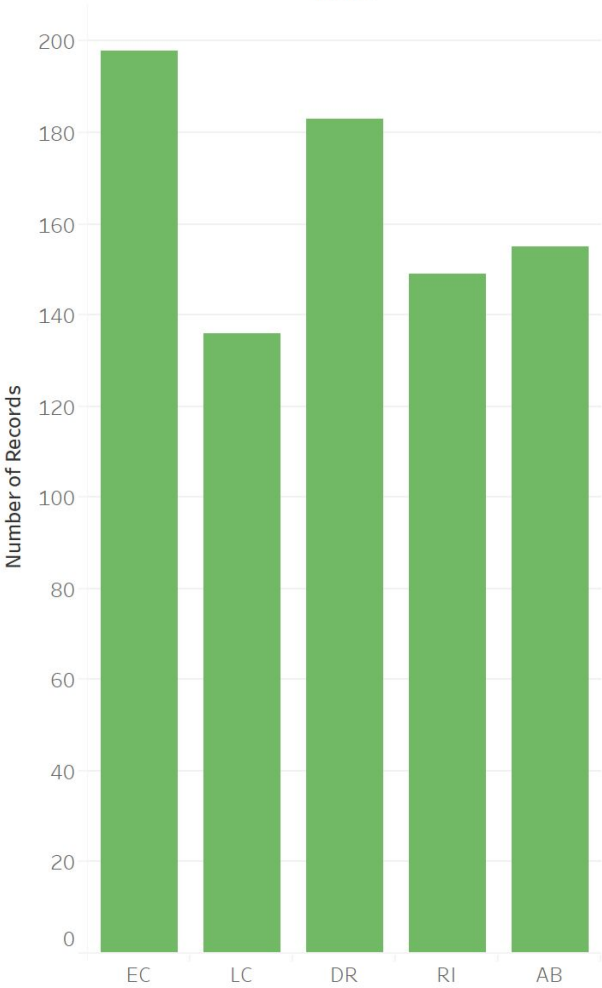
Color shows whether that feature was high or low for the dataset

Horizontal location shows whether the effect of that value caused a higher or lower prediction.

Updated EDA and Visualization Dashboard/Shiny

Tableau

File: /EDA/CompositionExploration.twb

Selected Findings & Hypotheses	Visualization												
<p><u>Finding 1:</u> From a macro-level, EC and DR have most observations and LC has the fewest observations</p>	<p>Number of observations for each Phase.</p> <p>Phase</p>  <table><thead><tr><th>Phase</th><th>Number of Records</th></tr></thead><tbody><tr><td>EC</td><td>198</td></tr><tr><td>LC</td><td>136</td></tr><tr><td>DR</td><td>183</td></tr><tr><td>RI</td><td>149</td></tr><tr><td>AB</td><td>155</td></tr></tbody></table>	Phase	Number of Records	EC	198	LC	136	DR	183	RI	149	AB	155
Phase	Number of Records												
EC	198												
LC	136												
DR	183												
RI	149												
AB	155												

## Finding 2:

For most species, their populations **shrink during LC** and **surge during DR**, while some species show reverse patterns.

What are the reasons?

## Hypothesis 1:

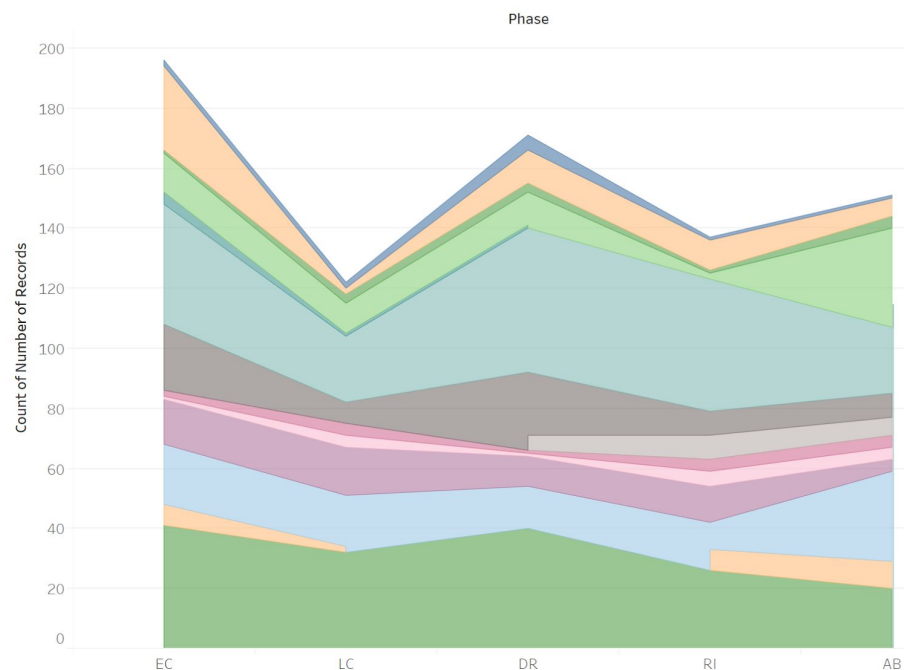
Number of observations in general is negatively affected by human activities

- **DR** has the **lowest** average level of human activities, and **LC** has the **highest** level [shown in the second graph]

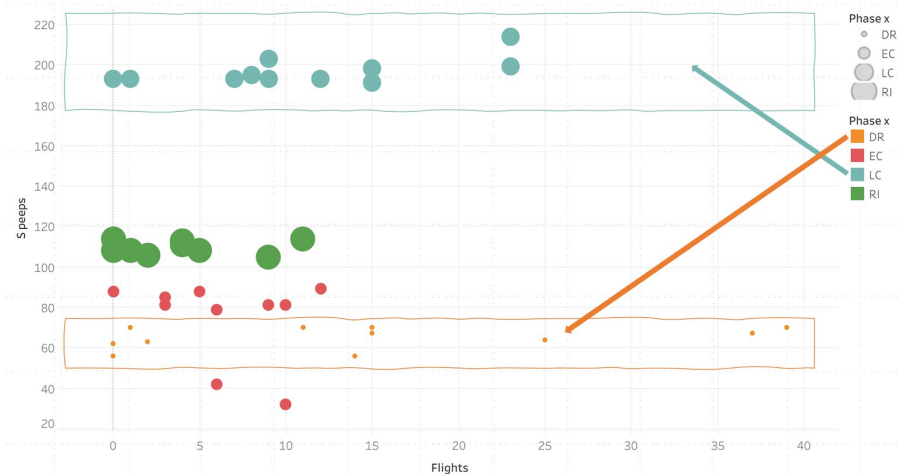
## Hypothesis 2:

Merely due to seasonality

Chg in # across Phases by Species (>10)



Flights vs Number of Workers



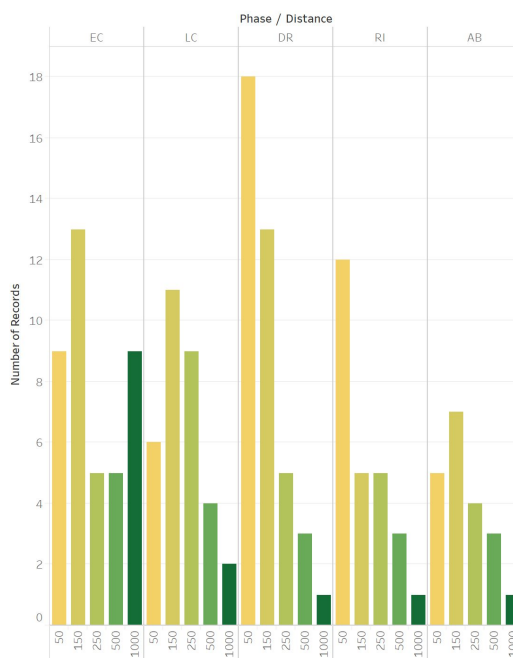
### Finding 3:

Species respond differently towards each phase. For example,

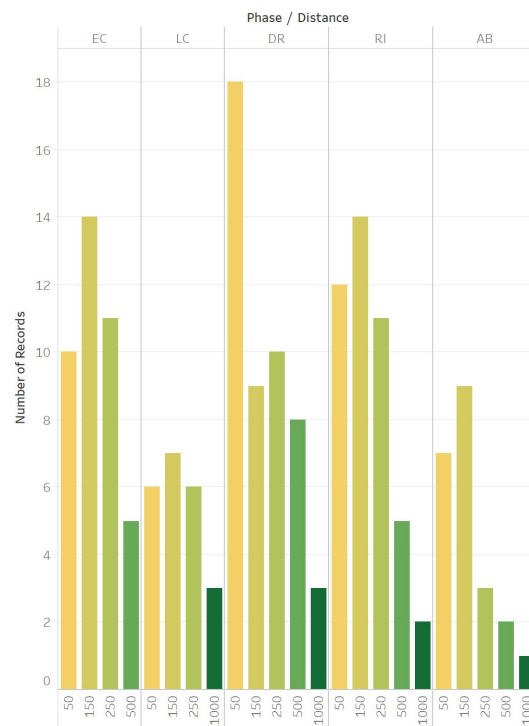
- In general, **more Danae and Tacana** were observed at plots **close to the construction site**.
- More Ockendeni were observed at plots **distant from the construction site**.

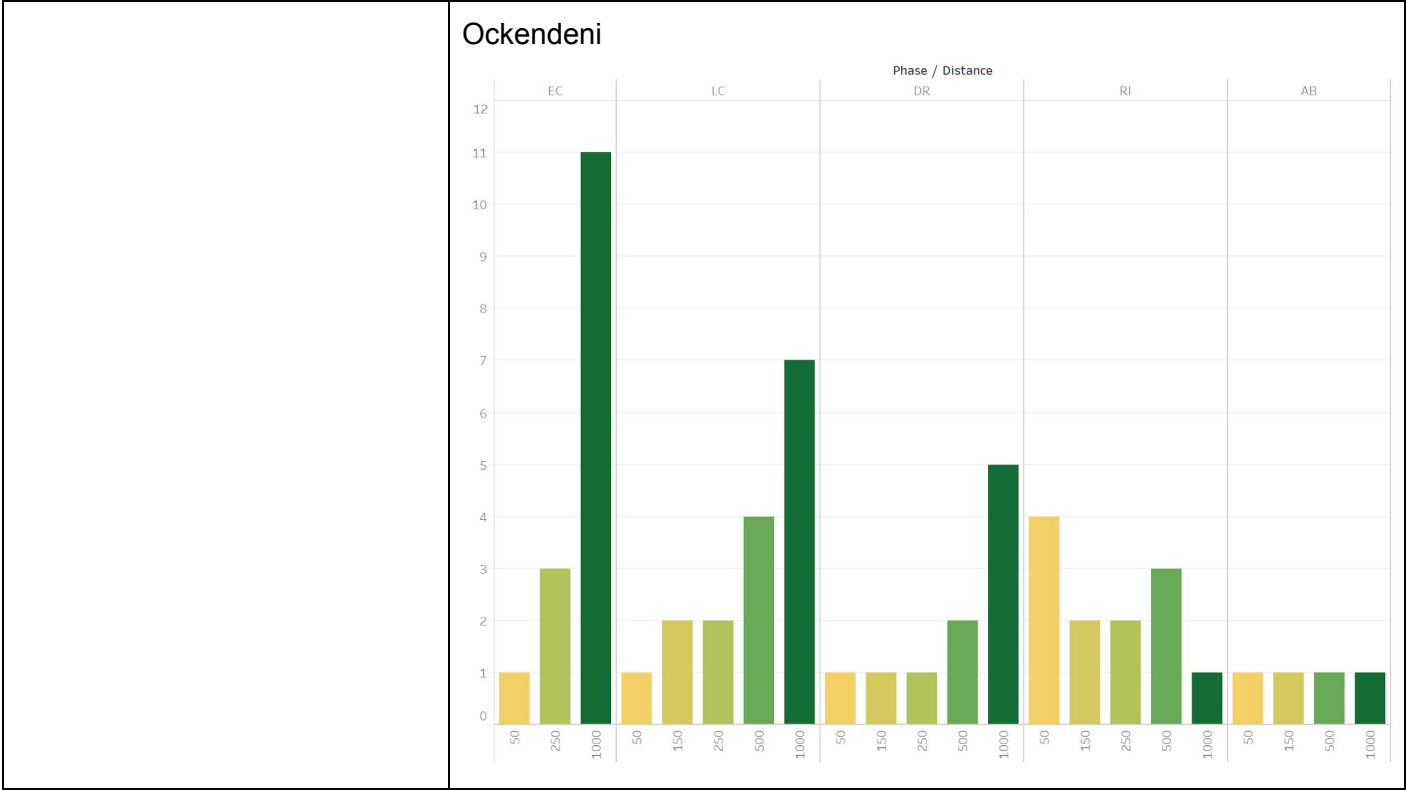
Is this due to change of habitat by the construction, or it is simply the original geography that makes the difference?

### Tacana



### Danae

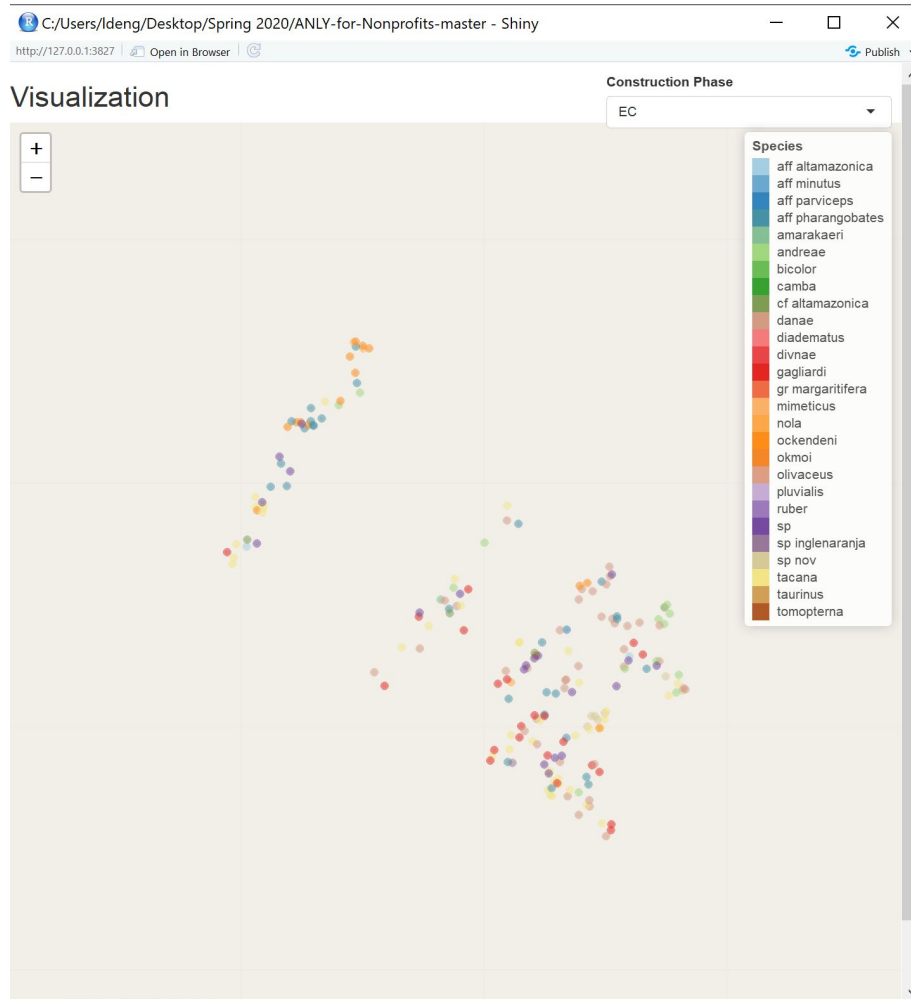




## Shiny

### 1. Map Visualization (Leaflet integrated Shiny App)

- The goal of the visualization is to show different composition of species at each phase & distance.
- The Scattered location for each observation is a randomly generated point within 30 meters of the plot where the observation is collected. The resulting dataset is `vis_data_v1.csv`.
- A dropdown bar is designed for users to filter different construction phases. Each point on the map is an observation. The species of each observation is encoded by a color.
- One drawback of this vis is the difficulty to distinguish species since 27 colors are used to encode the 27 species. Fewer insights can be derived by simply looking at this vis.
- Code file: `map_app.R`

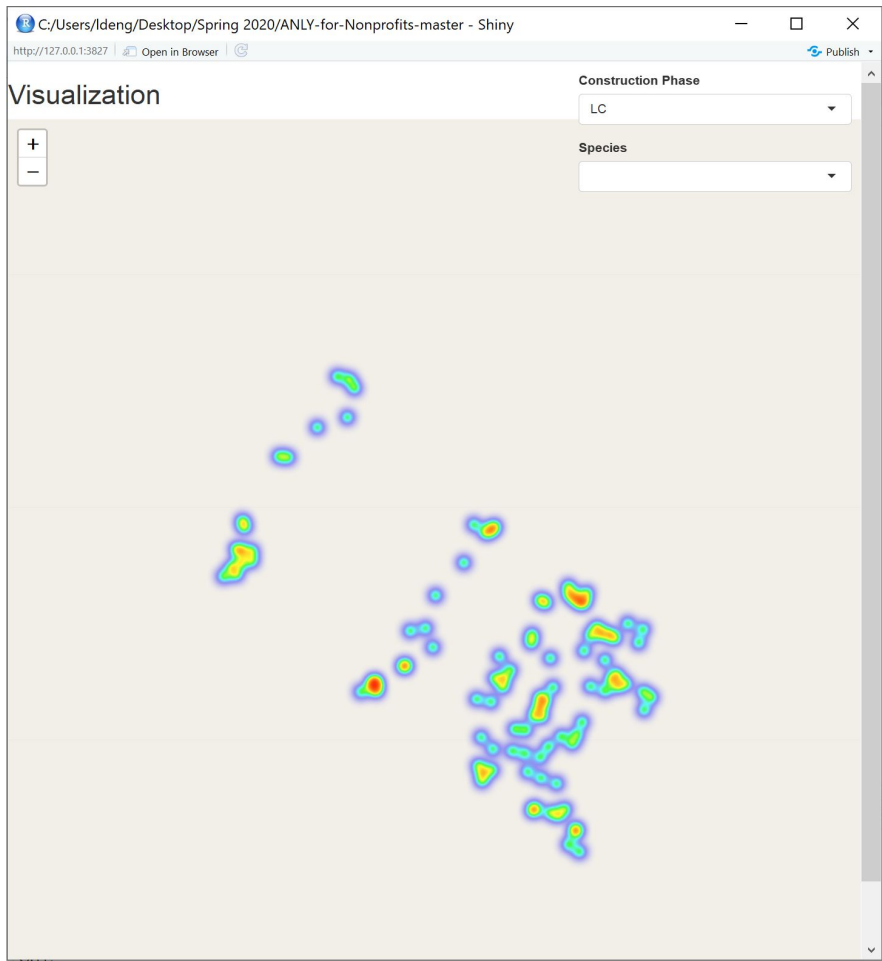


## 2. Geo Heat Map Visualization

- The goal of this geo heat map visualization is to show the geo distribution of each species at each phase and distance
- Two dropdown bars are designed for users to filter different construction phases and species (the species one is still under development). The density of observations is encoded by a color scale from blue (less concentrated) to red (more concentrated).
- Code file: heatmap\_app.R



Amphibian Response to Natural Gas Development in Puruvian Amazon Forest



## Plotly

### 1. Body Mass/Body Length ratio by Phase

- Since we are wondering if the species compositions and properties were changed in different phases, we plotted the amphibians composition by phase. The data points are colored by genus and the size is the ratio of mass divided by body length. Here is the link for the interactive graph: <https://plot.ly/~yyyyyokoko/122/#/>
- One interesting finding is that the species have a smaller mass ratio during drilling phase comparing to early construction.

