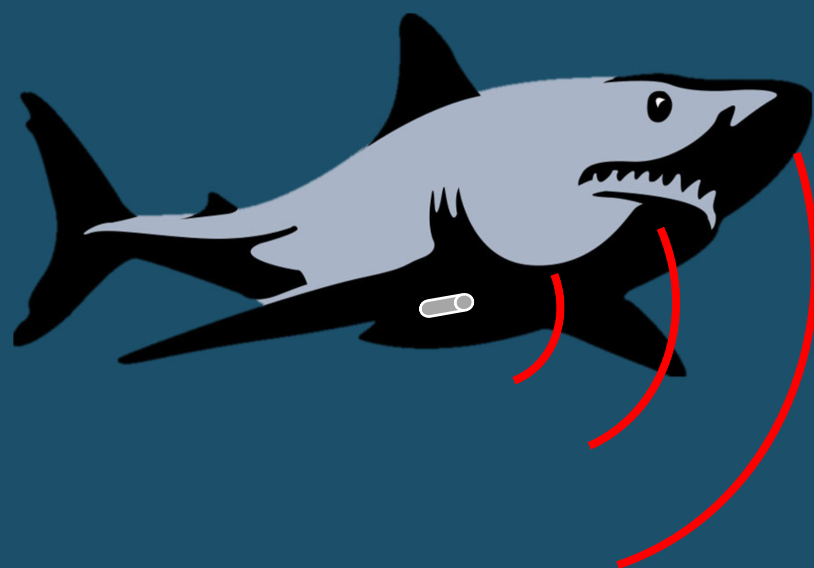
Photo by Ralph Pace

Using Acoustic Telemetry Data and Environmental Variables to Predict White Shark Presence in Southern California

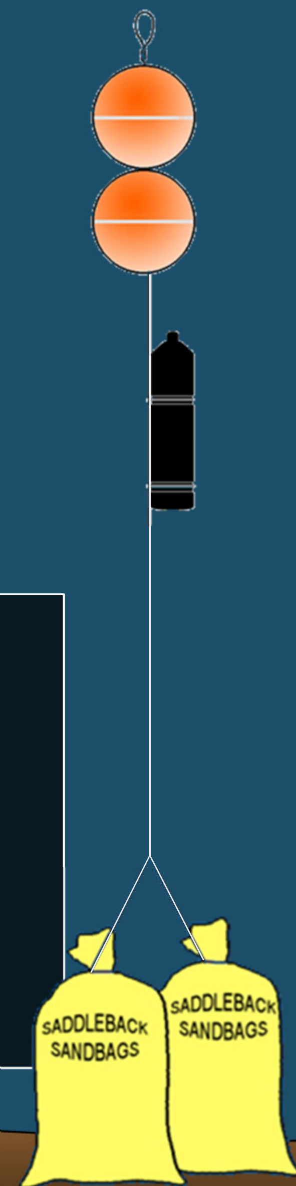
Echelle Burns – Jan 2020

Introduction: Acoustic Telemetry



Common tool to study animal movement

- Transmitters emit ultrasonic sound
- Acoustic receiver picks up transmitter signal
  - Records: Date, Time, Transmitter Code, Mooring Location



Introduction: Study Species

Juvenile White Sharks (*Carcharodon carcharias)*

- Apex predator with a global distribution
- Repeated movements to nearshore waters off southern California
  - Seasonal/annual scale

Running hypothesis: Temperature Driven?

Introduction: Project Objectives

Objectives:

- Determine which combinations of environmental parameters influence White Shark presence in southern CA
- Predict ranges of environmental conditions that are most likely to result in high densities of White Sharks

Audience:

- Researchers using similar species or technology
- Lifeguards and public safety personnel

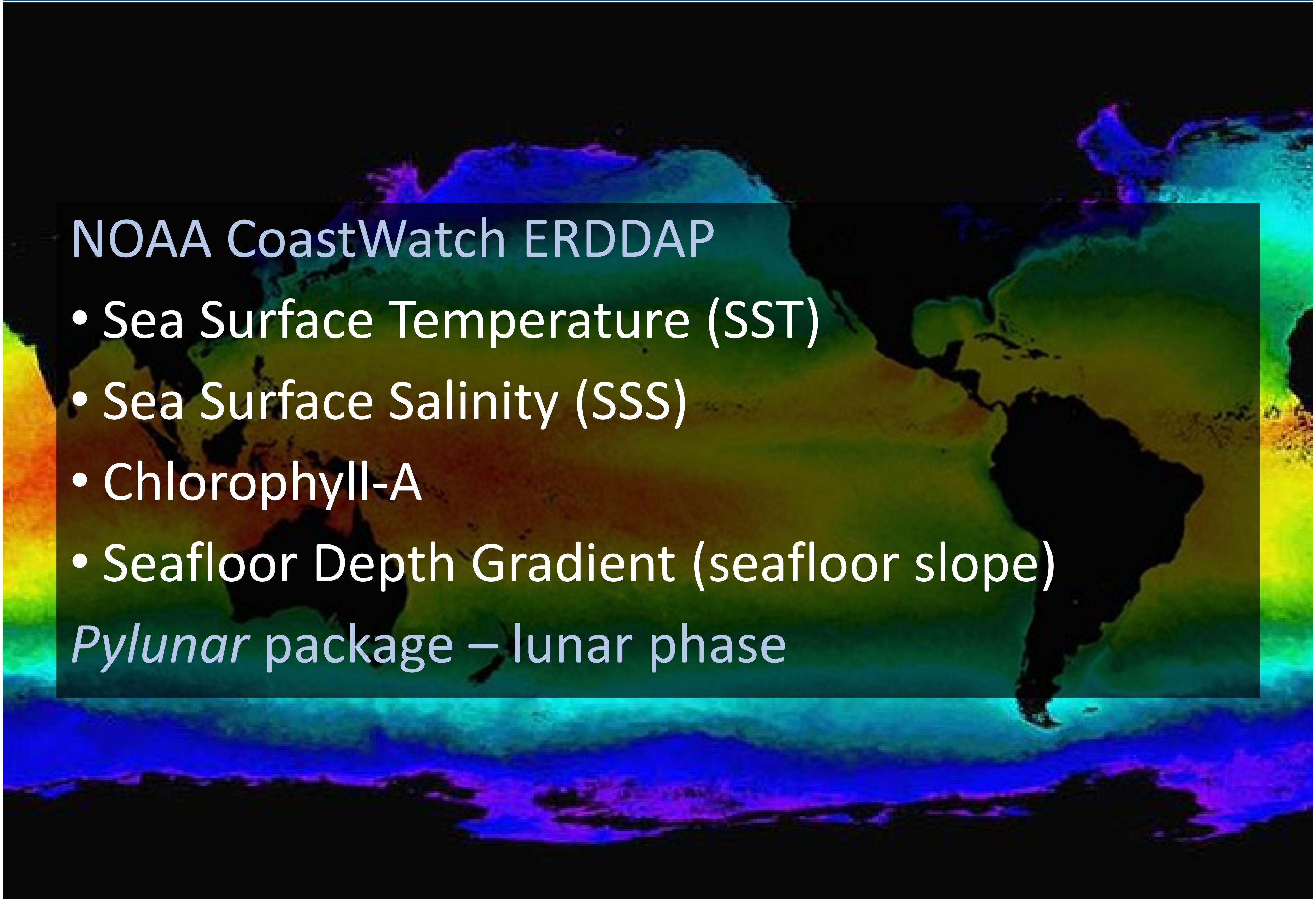
Materials and Methods: Data Sources

Cal State University, Long Beach Shark Lab:

- Receiver Deployment Log
  - When and where receivers were deployed

- Transmitter Deployment Log
  - Tagging date, location, species
- Receiver Data (2012 – 2019)
  - Detection data from nearby animals

Materials and Methods: Data Sources



NOAA CoastWatch ERDDAP

- Sea Surface Temperature (SST)
- Sea Surface Salinity (SSS)
- Chlorophyll-A
- Seafloor Depth Gradient (seafloor slope)

Pylunar package – lunar phase

Materials and Methods: Data Cleaning

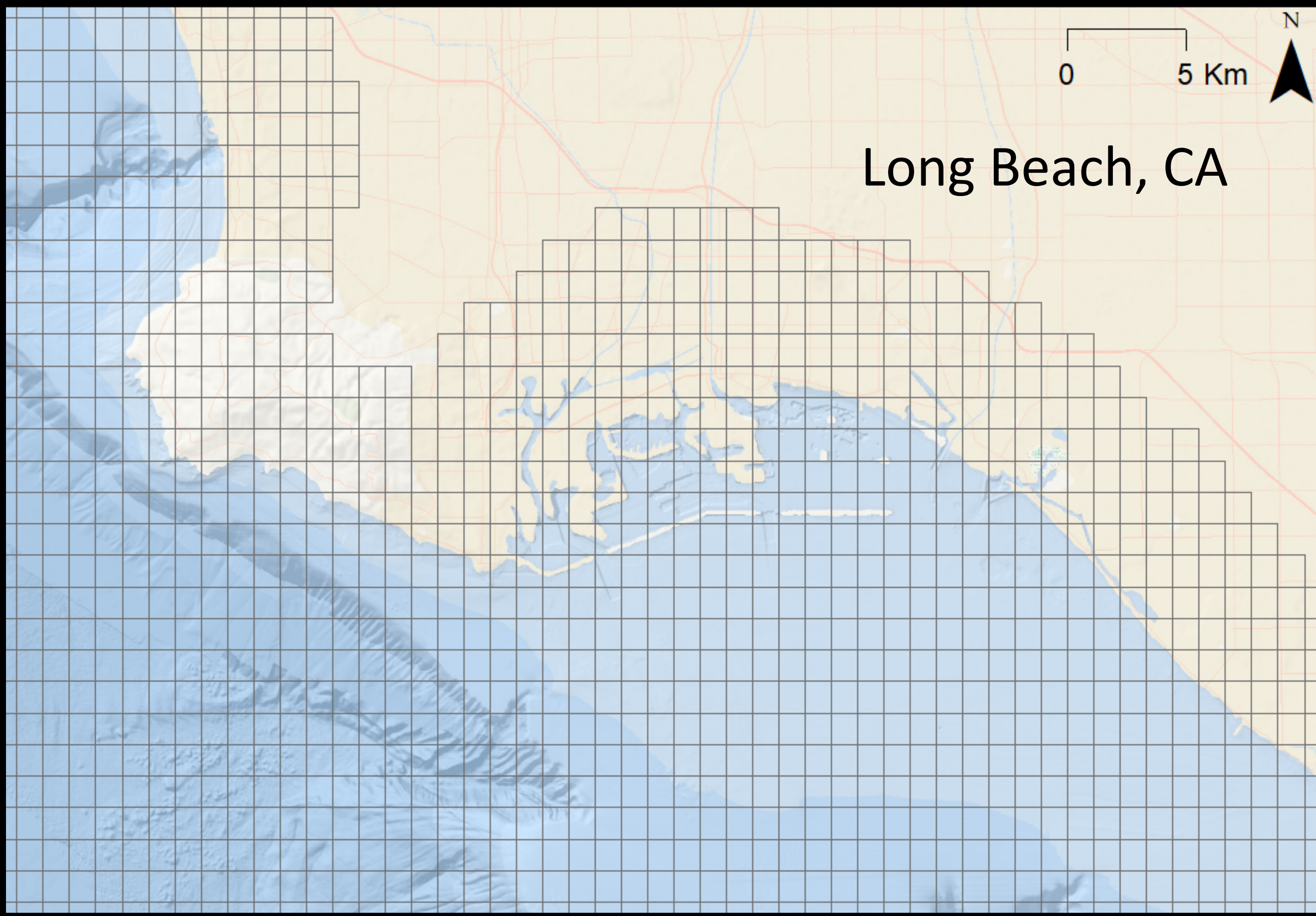
Acoustic Receiver Data

- White Sharks (WS) only
- Individual heard 2+ times per day at a receiver
- Receiver confirmed to have been in the water
- Latitude/Longitude reflect receiver deployment location

Result: WS detection data from 2014 to 2019

Materials and Methods: Data Cleaning

Fishnet 0.01 x 0.01 decimal degree (ArcMap)



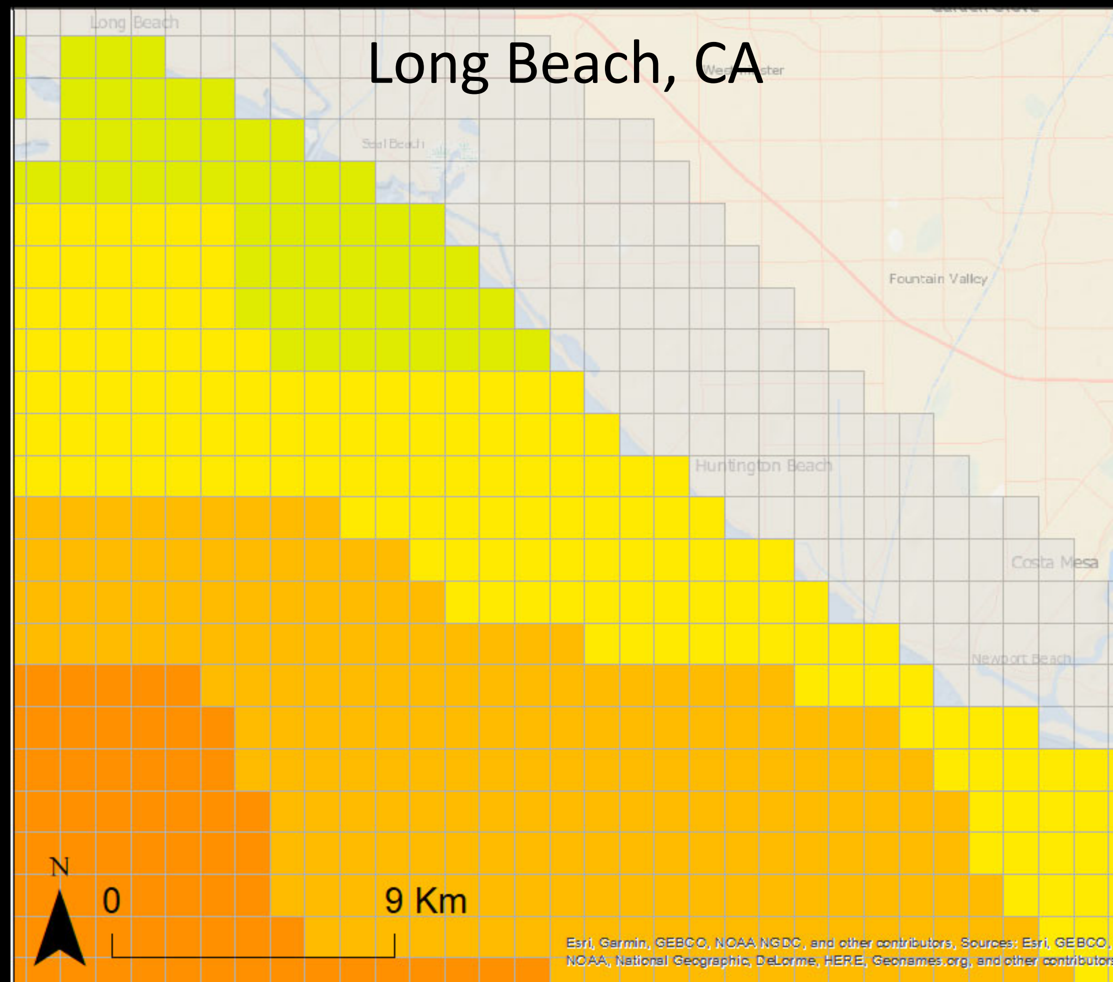
To merge environmental data with acoustic receiver data

Materials and Methods: Data Cleaning

Fishnet 0.01 x 0.01 decimal degree (ArcMap)

Nearest Neighbor for cells too close to shore

Mean value if > 1 measurements per cell

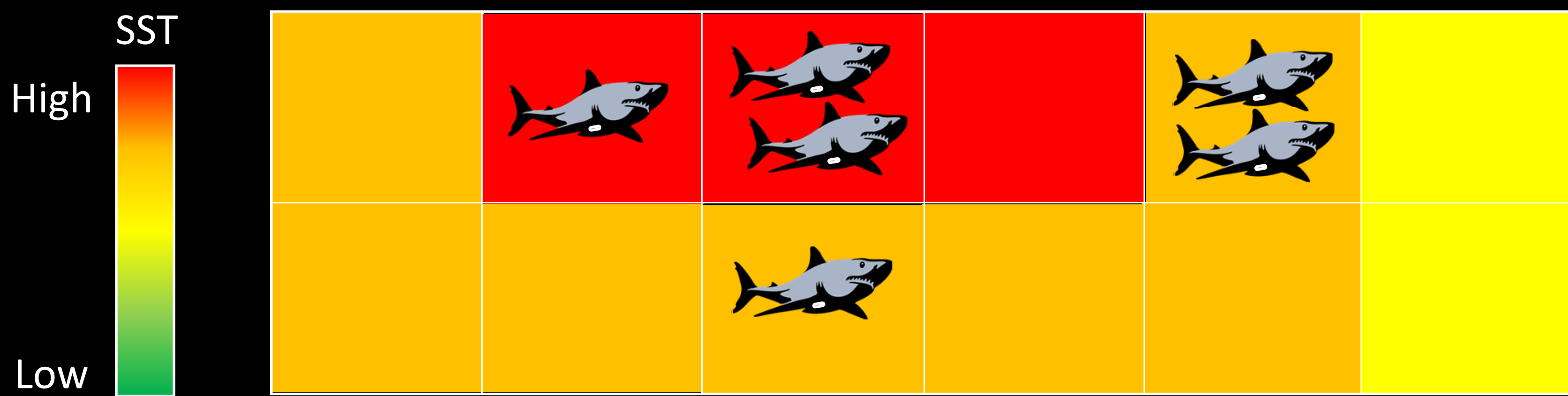


To merge environmental data with acoustic receiver data

Materials and Methods: Data Cleaning

Zero-Inflated Dataset

- Analyzed by year
- Kept all data where sharks were present



- Jackknifed data for when sharks were absent
  - *Method 1*: randomly sample nearby Zones (included times when receiver density = 0)
  - *Method 2*: grab all samples where receiver density > 0 and up-sample until categories were balanced

Materials and Methods: Data Cleaning

Individual shark data changed to # unique sharks per zone per day (Shark Density)

Final Dataset

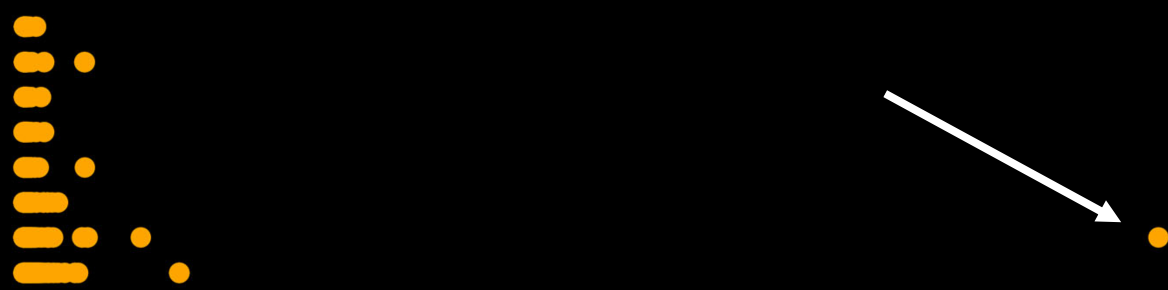
Response: *Shark Density*

Categorical Predictors:	***Continuous Predictors:***
Month	*SST*
Year	*SSS*
Zone ID	*Chlorophyll-A*
Receiver Density	*Seafloor depth gradient*
Lunar Phase	

Materials and Methods

Visual Assessments:

- Data visually assessed for outliers

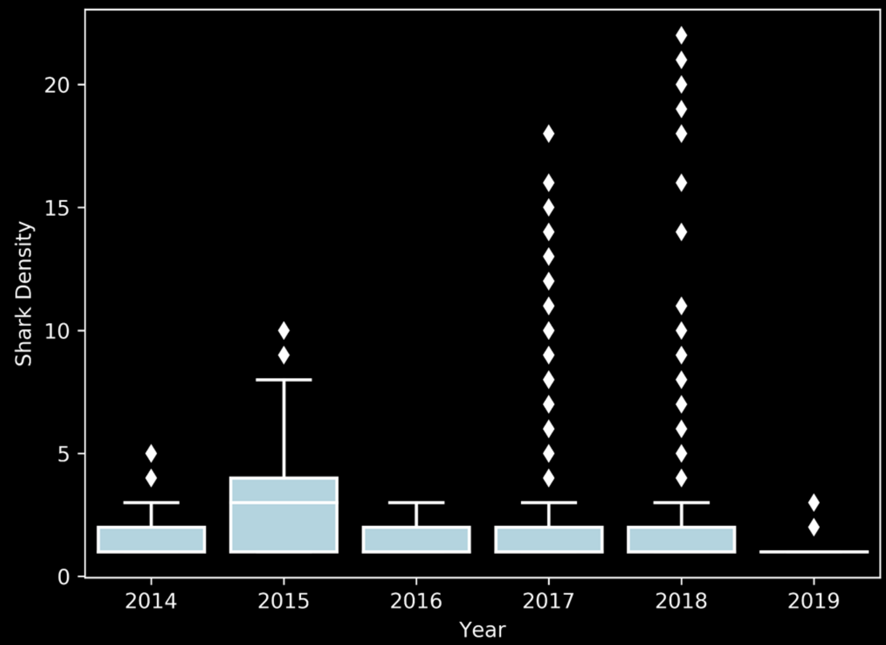


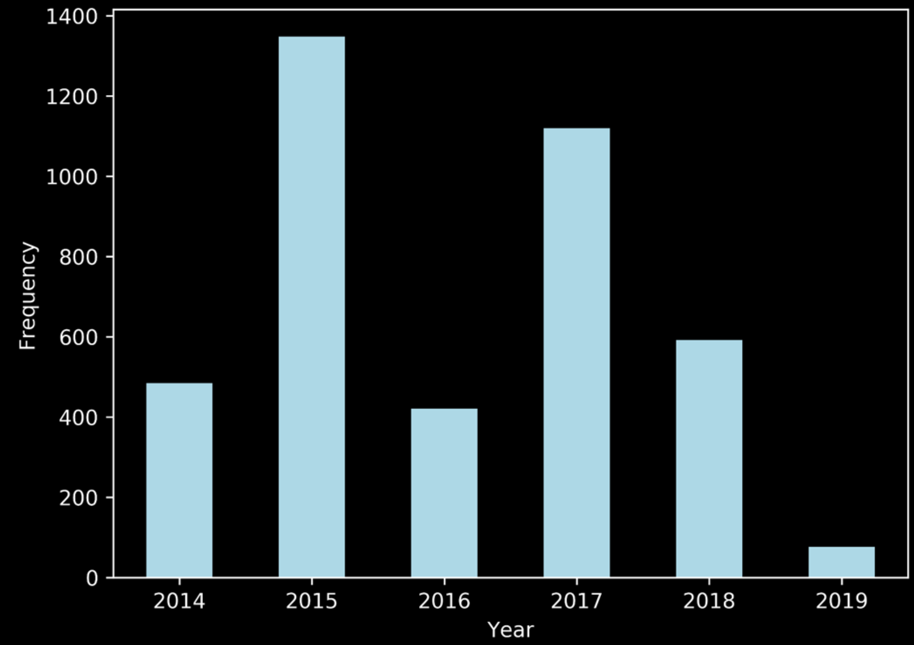
Statistical Analyses:

- Ordinary Least Squares ANOVA (Receiver Density, Year, Zone ID)
  - Response variable: shark density
  - Only when shark presence > 0
- T-Tests (SST, SSS, Chlorophyll-A, depth gradient)
  - Response variable: shark presence vs absence

Results: Shark Density v Year

**Data in this analysis used method 1 for jackknifing*





Ordinary Least Squares ANOVA

Shark Density across years compared to 2014

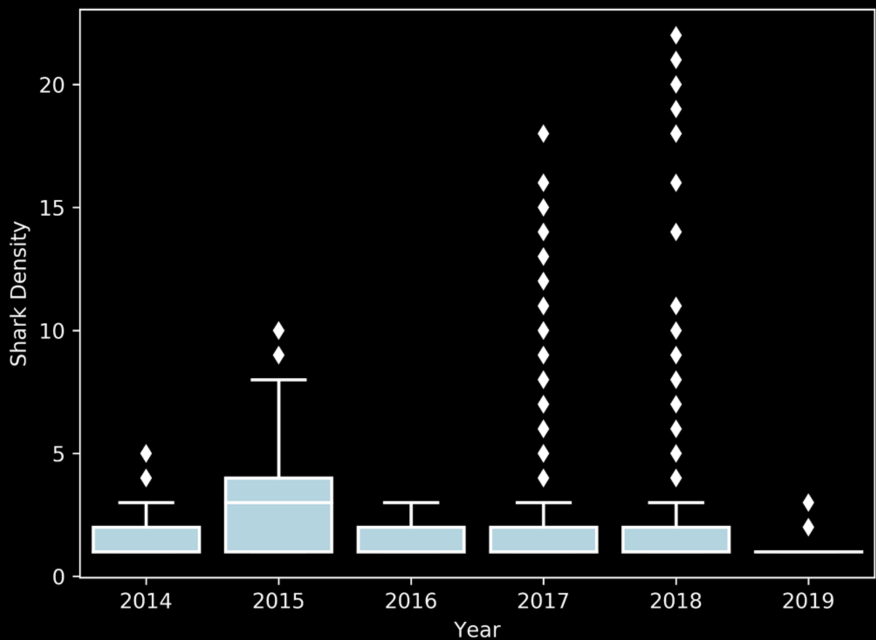
2015, 2016, 2017 >>> 2014 ($p < 0.001$)

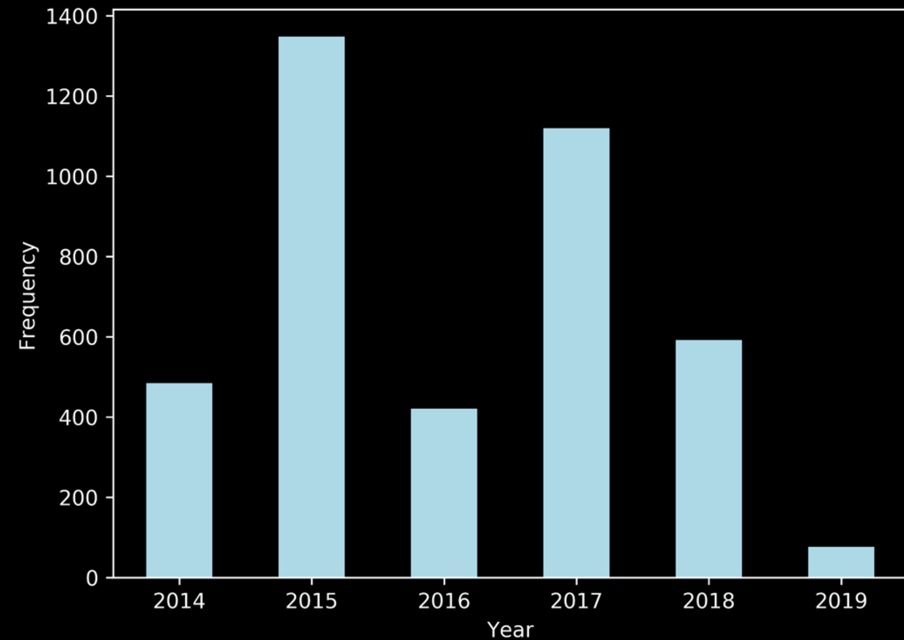
2017 == 2014 ($p = 0.153$)

2019 <<< 2014 ($p = 0.028$)

Discussion: Shark Density v Year

**Data in this analysis used method 1 for jackknifing*





2015, 2017, 2018 >>> 2014 ($p < 0.001$)

2016 == 2014 ($p = 0.153$)

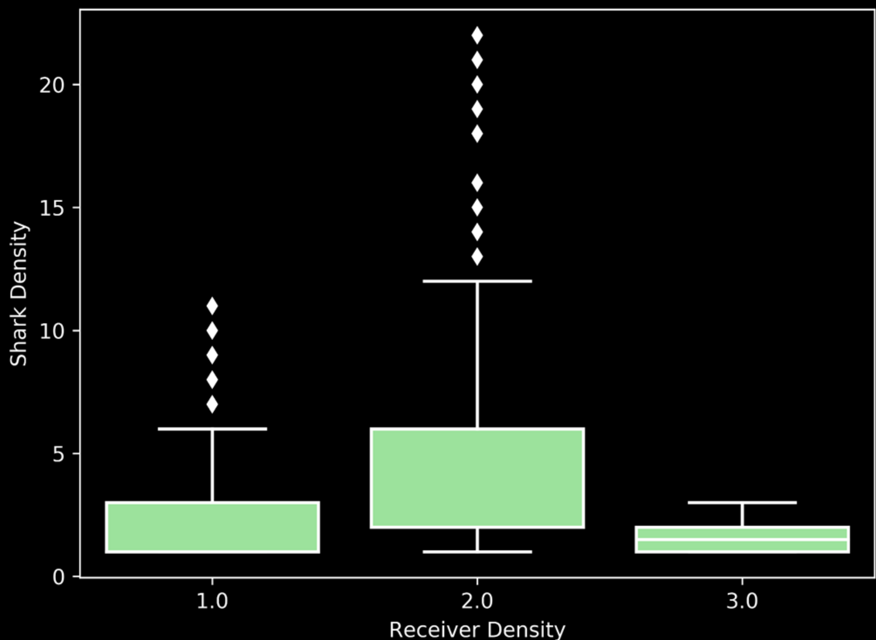
2019 <<< 2014 ($p = 0.028$)

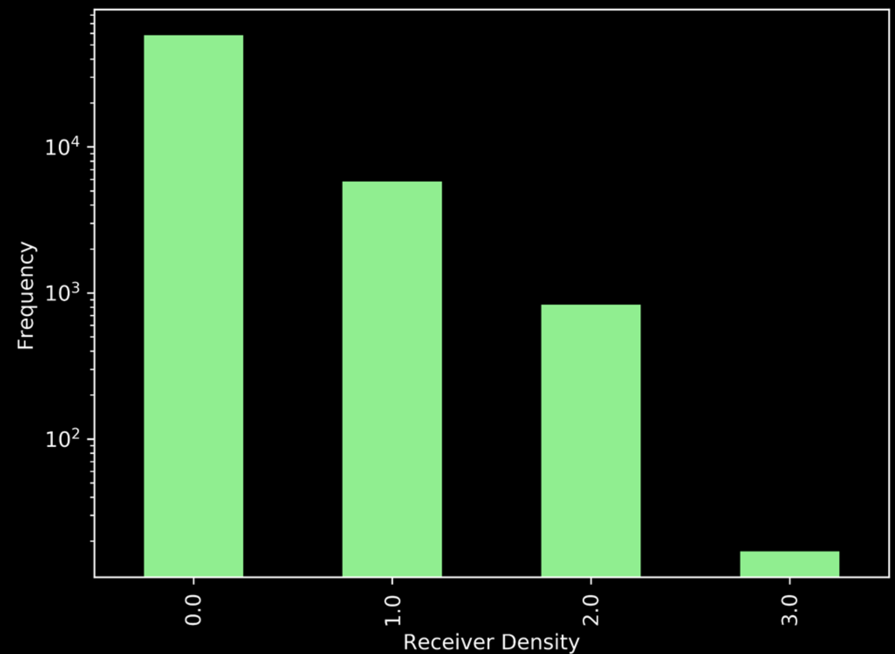
More ideal environmental conditions in 2015, 2017, 2018?

Fewer data in 2019 because the year was not yet over?

Results: Shark Density v Receiver Density

**Data in this analysis used method 1 for jackknifing*





Ordinary Least Squares ANOVA

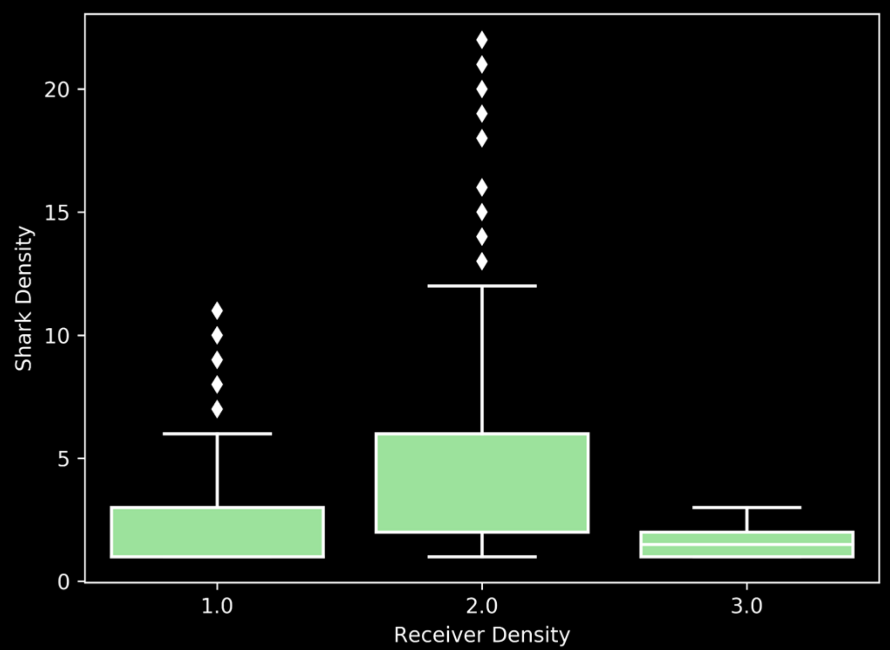
Shark Density across receiver densities

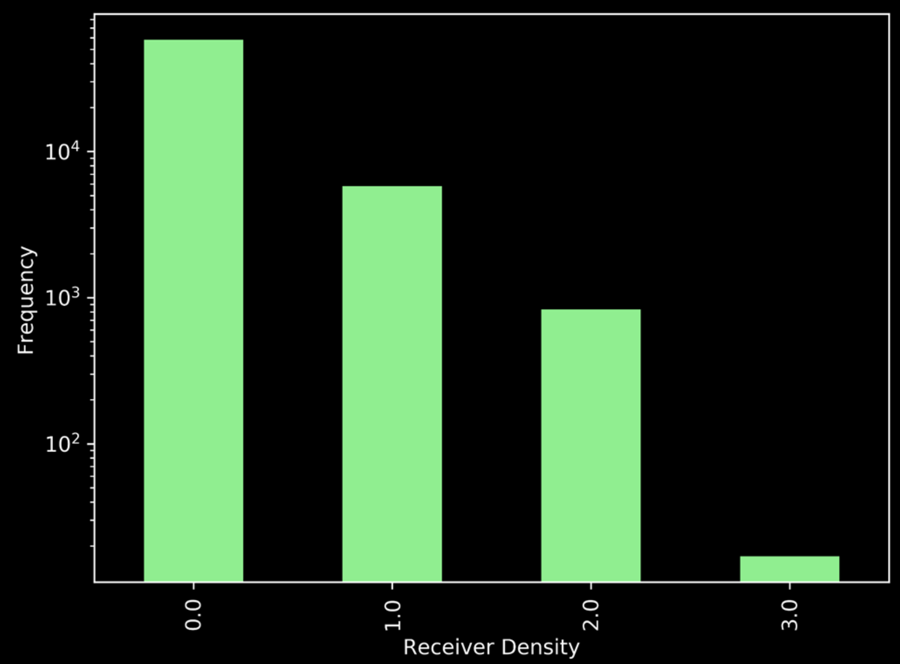
2 receivers >>> 1 receiver ($p < 0.001$)

3 receivers == 1 receiver ($p = 0.625$)

Discussion: Shark Density v Receiver Density

**Data in this analysis used method 1 for jackknifing*





2 receivers >>> 1 receiver ($p < 0.001$)

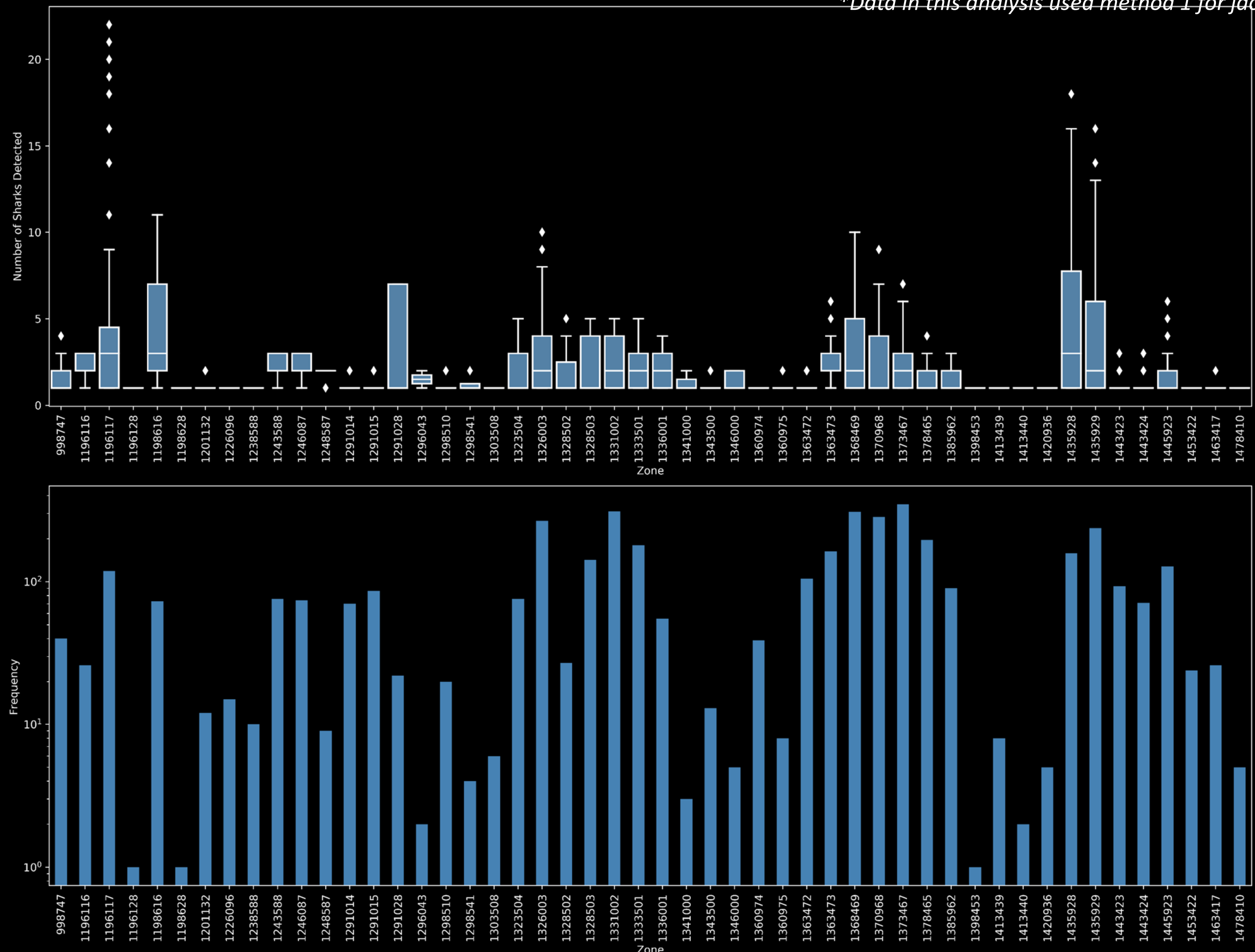
3 receivers == 1 receiver ($p = 0.625$)

More receivers in a grid cell = more cell coverage

Few instances in which receiver density > 2 receivers

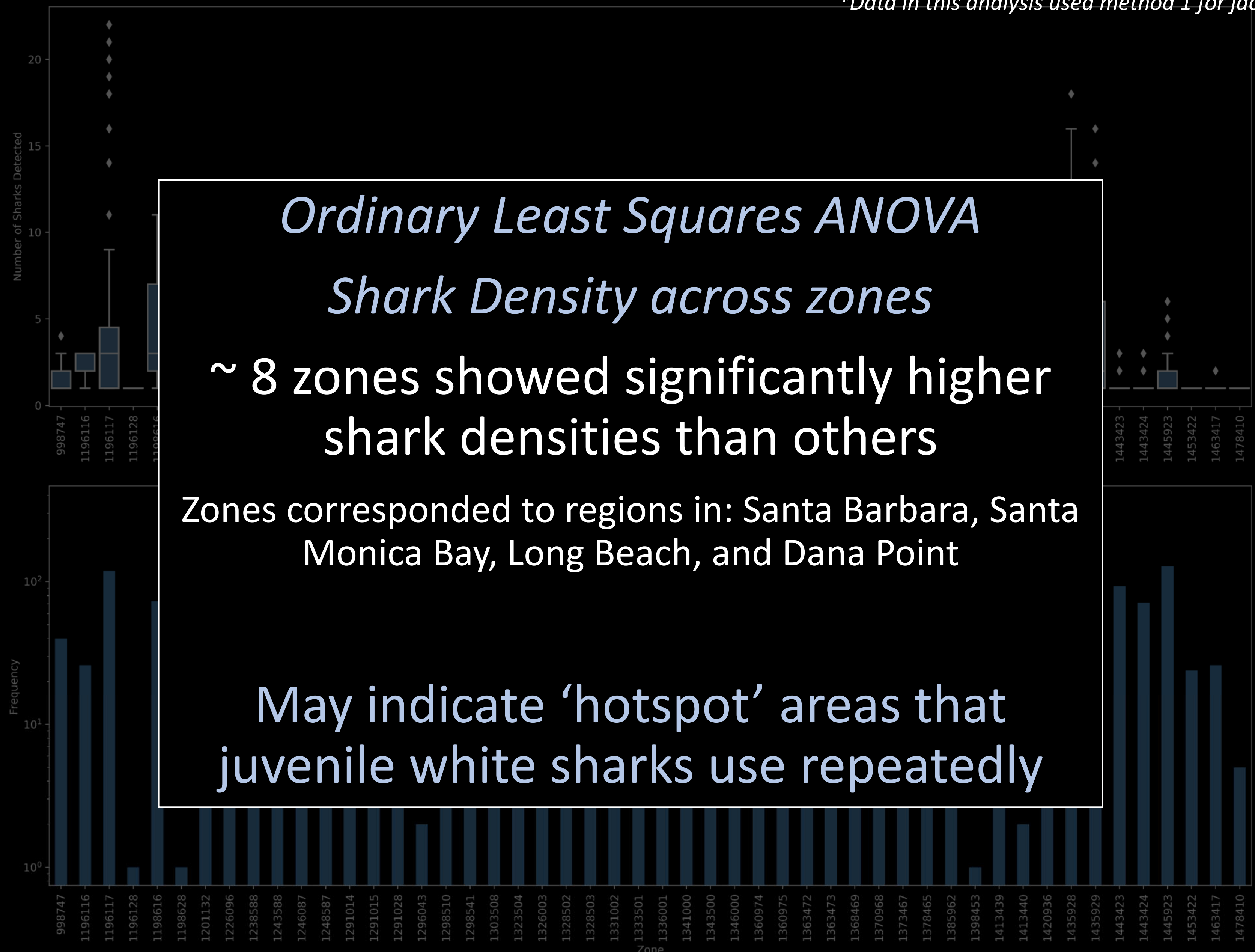
Results: Shark Density v Zone

**Data in this analysis used method 1 for jackknifing*



Results and Discussion: Shark Density v Zone

**Data in this analysis used method 1 for jackknifing*



Ordinary Least Squares ANOVA

Shark Density across zones

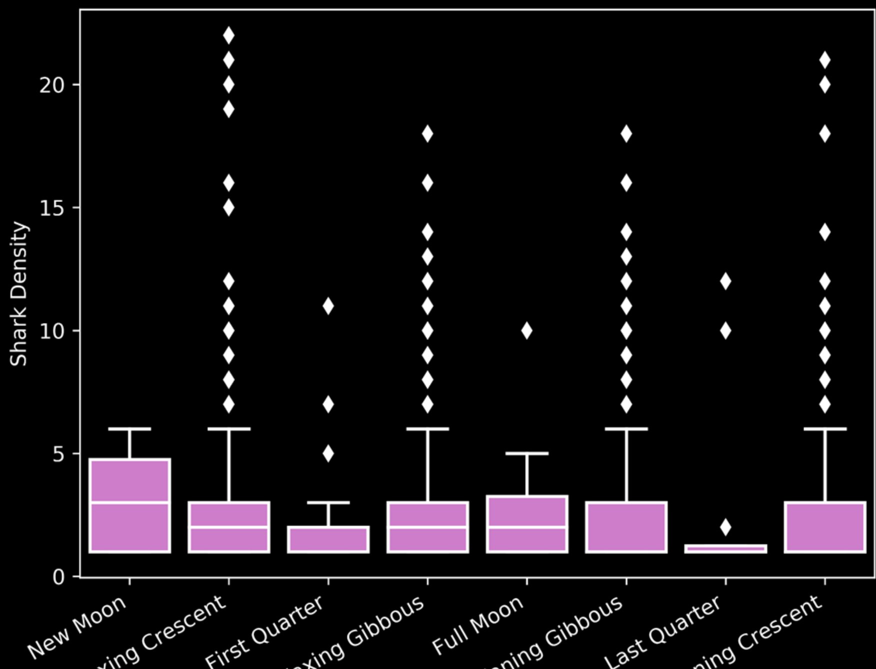
~ 8 zones showed significantly higher shark densities than others

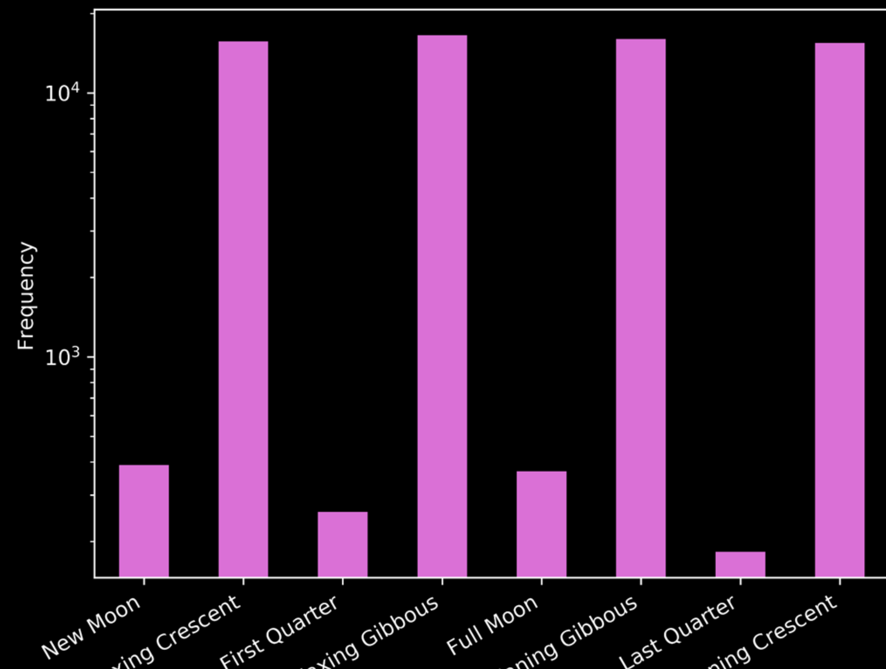
Zones corresponded to regions in: Santa Barbara, Santa Monica Bay, Long Beach, and Dana Point

May indicate 'hotspot' areas that juvenile white sharks use repeatedly

Results: Shark Density v Lunar Phase

**Data in this analysis used method 1 for jackknifing*





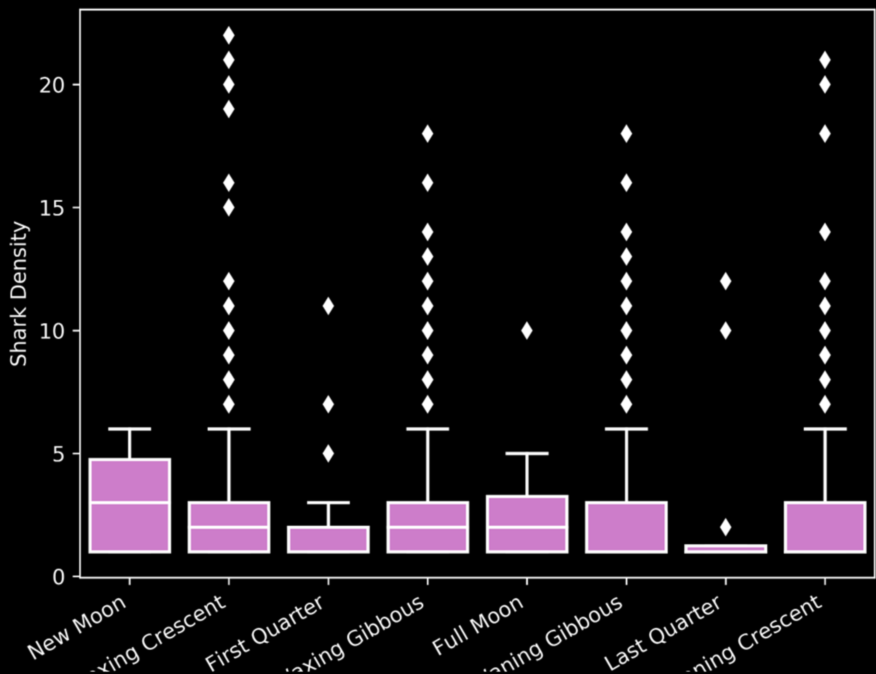
Ordinary Least Squares ANOVA

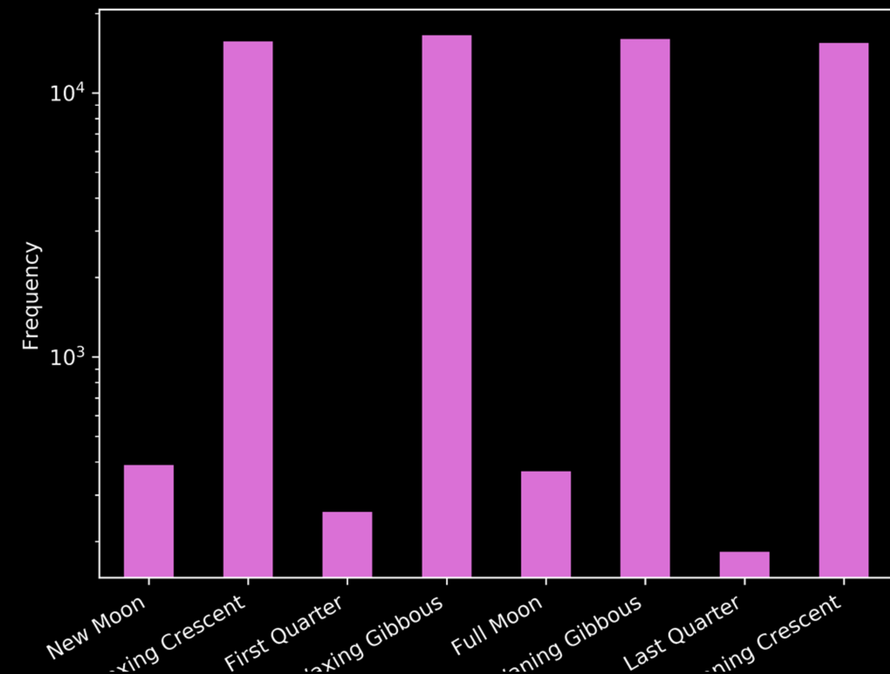
Shark Density across lunar phases

No significant difference ($p > 0.05$)

Discussion: Shark Density v Lunar Phase

**Data in this analysis used method 1 for jackknifing*



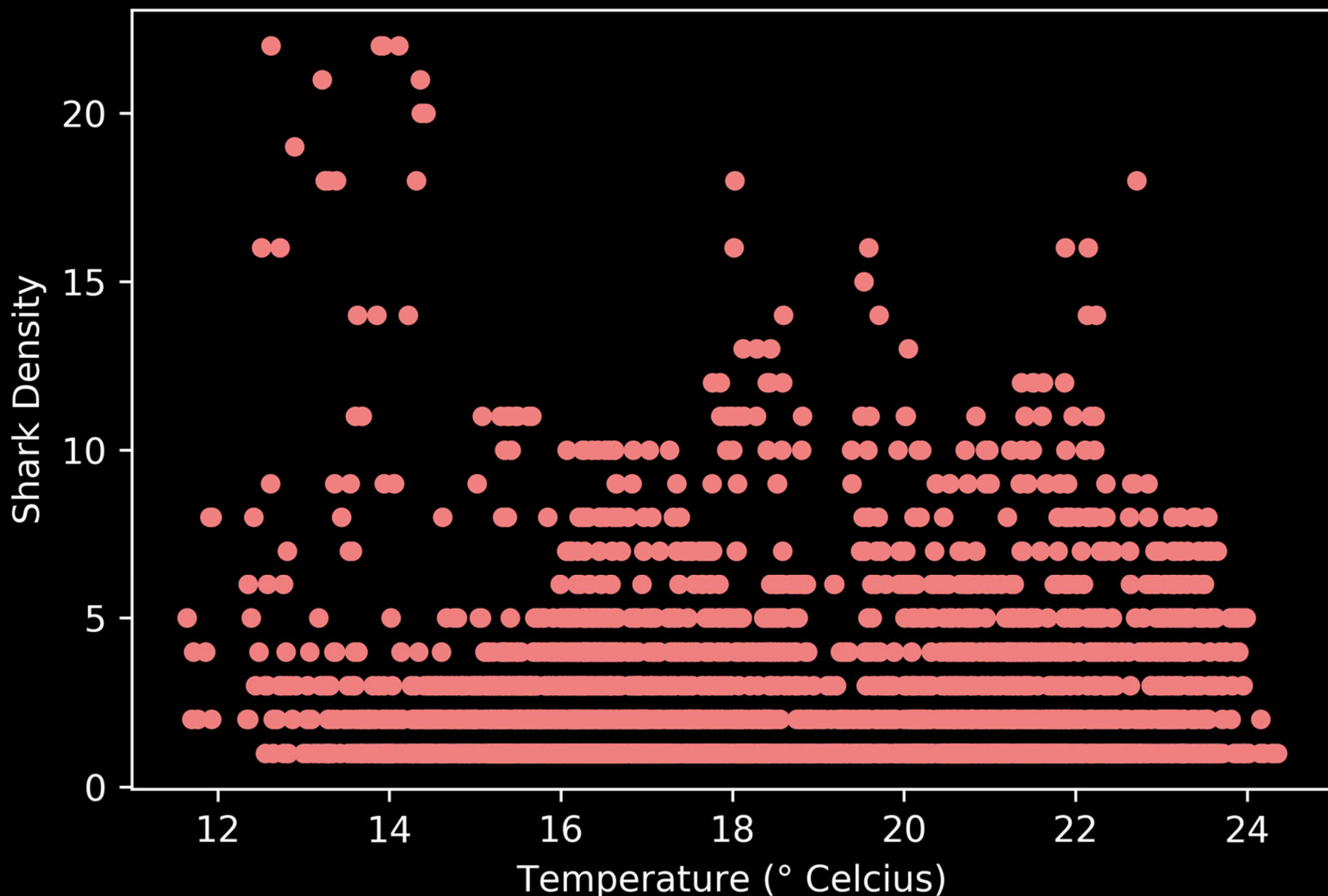


No significant difference ($p > 0.05$)

Juvenile White Shark movements are not driven by the lunar cycle

Results: Shark Density v Sea Surface Temperature

**Data in this analysis used method 1 for jackknifing*

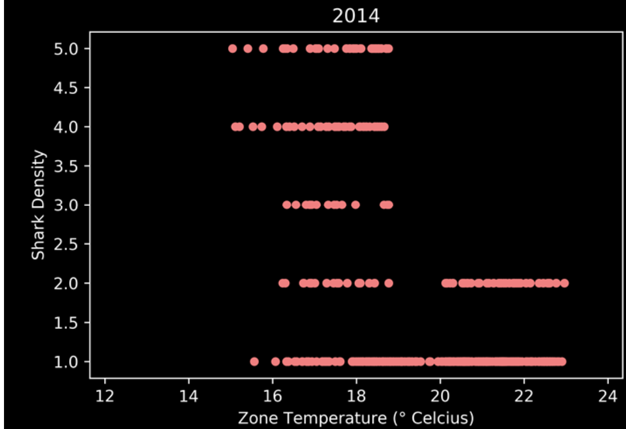


Min: 10.99
Max: 25
Median: 17.88
STD: 2.66

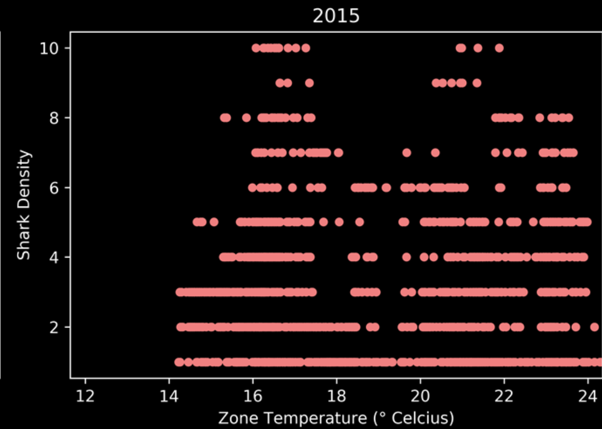
Highly variable shark densities at all temperature values

Results: Shark Density v Sea Surface Temperature

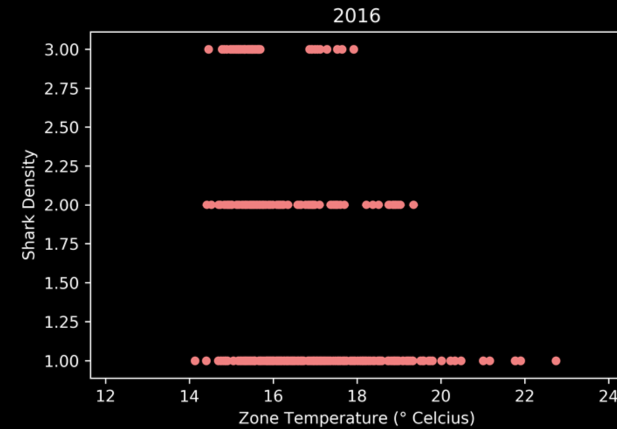
2014: 16-17^{o}C



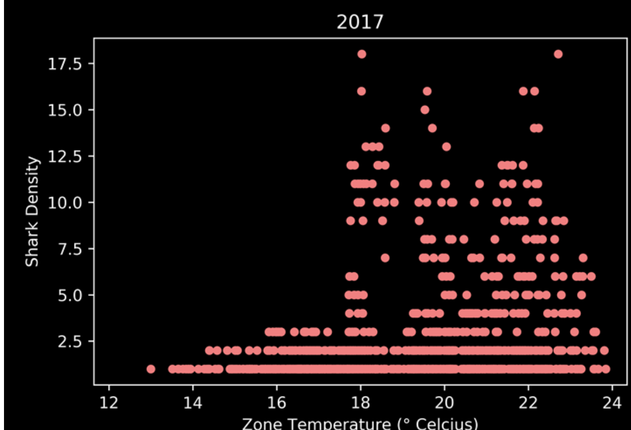
2015: 16-17^{o}C, 20-22^{o}C



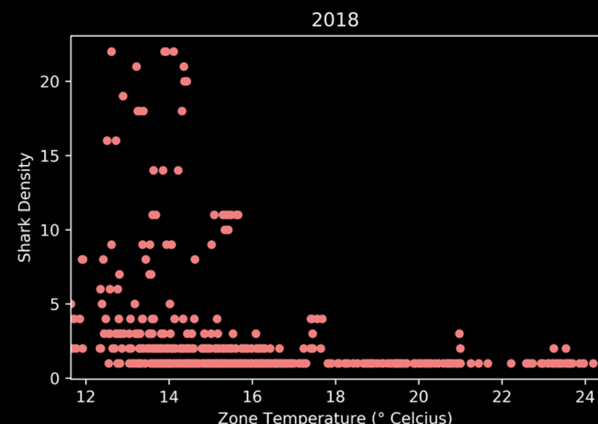
2016: 14-18^{o}C



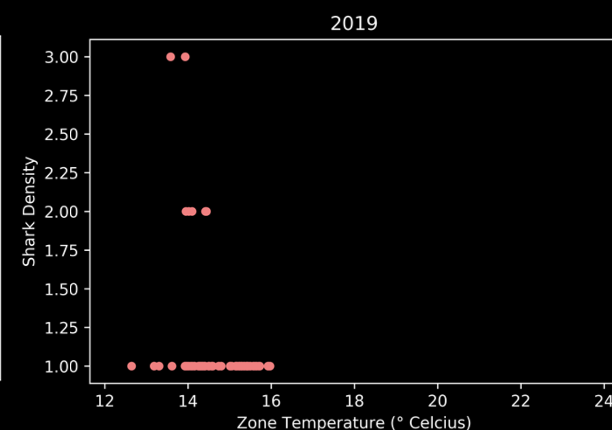
Different years show different relationships between SST and shark density.



2017: 18-22^{o}C



2018: 12-14^{o}C

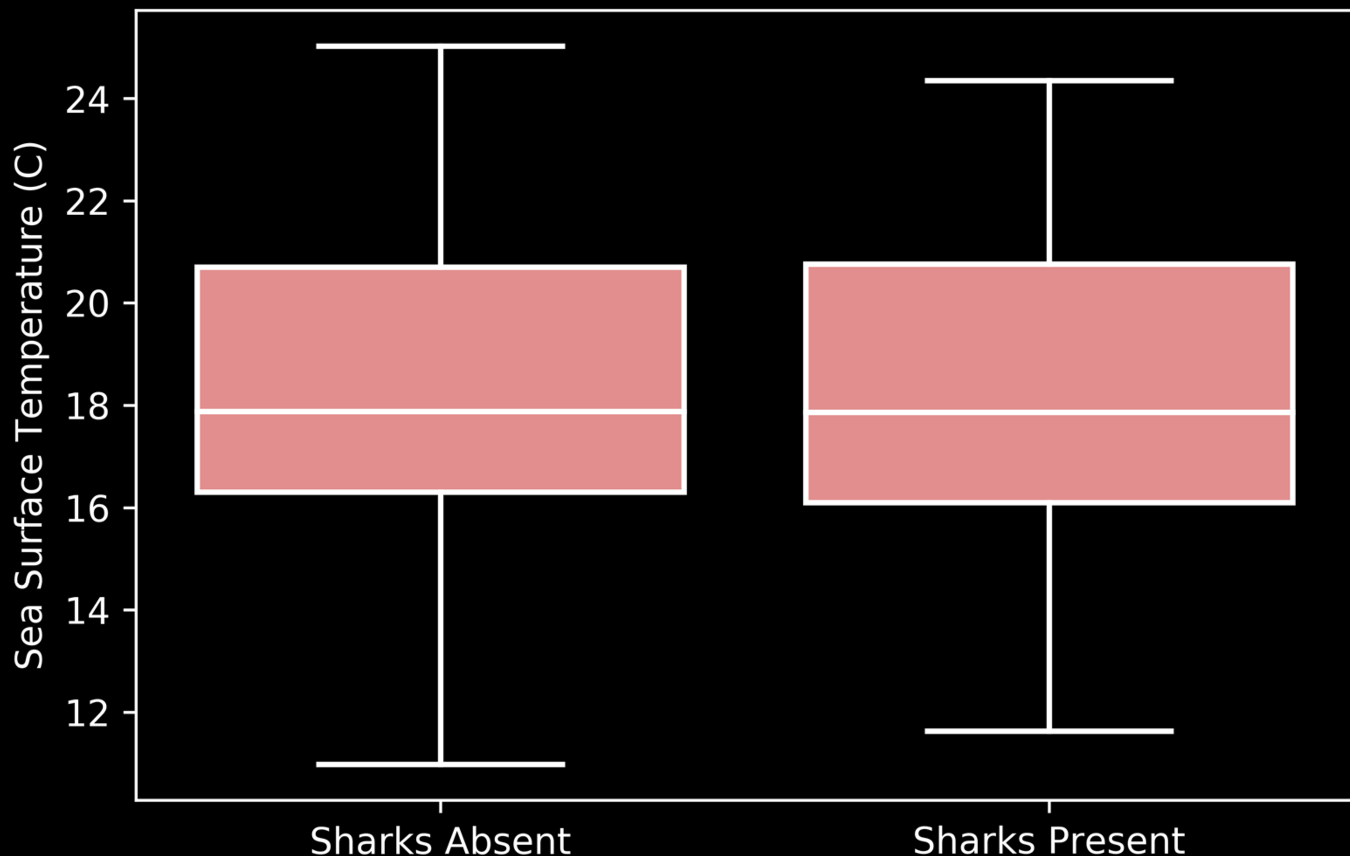


2019: 13-14^{o}C

**Data in this analysis used method 1 for jackknifing*

Results: Shark Presence v Sea Surface Temperature

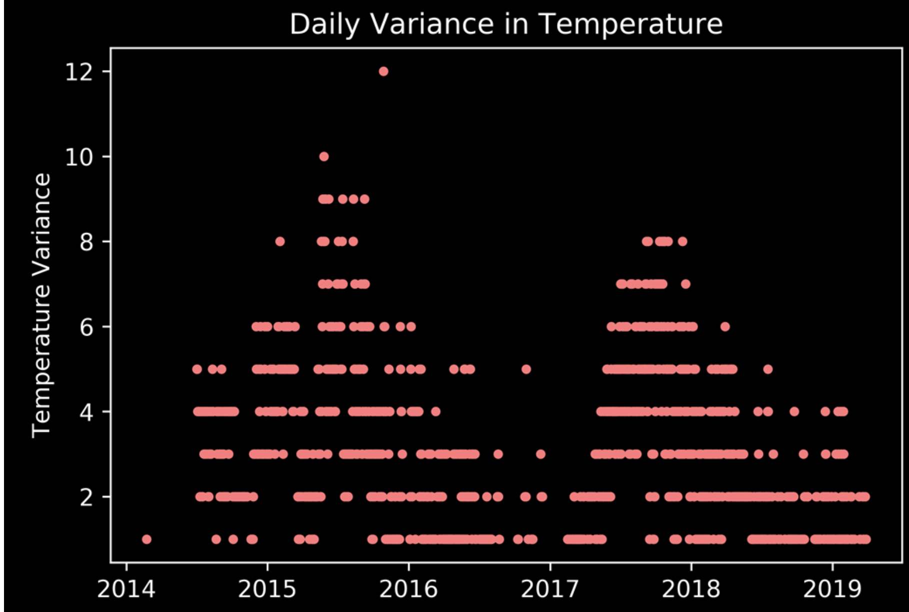
**Data in this analysis used method 1 for jackknifing*

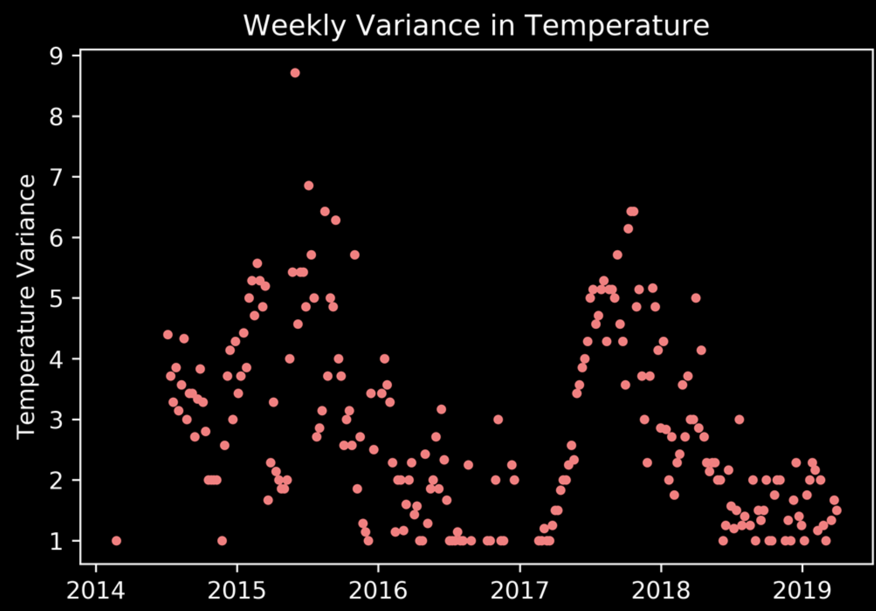


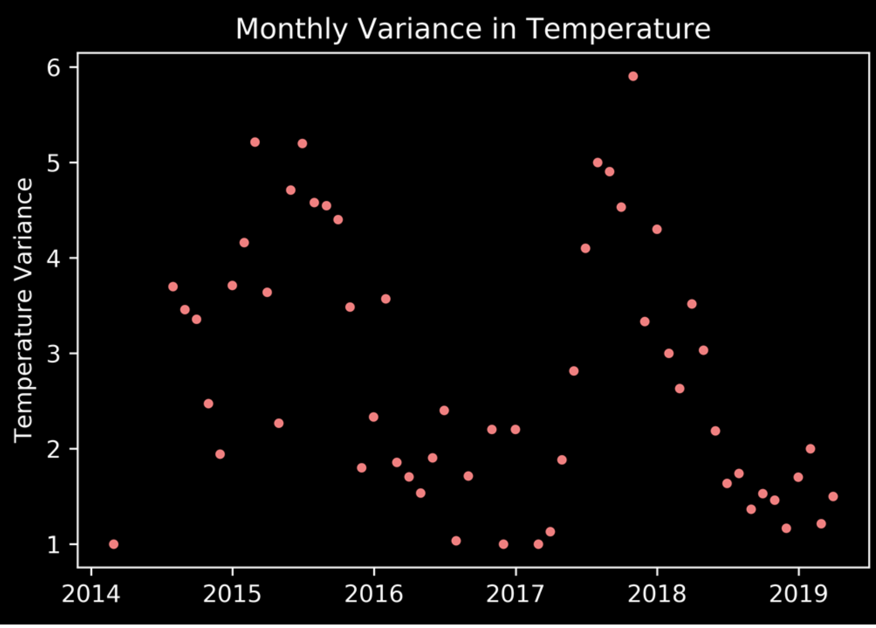
Mean (± SD): 18.39 ± 2.65°C   8.31 ± 2.81°C

No significant difference ($t = -1.824$, $p = 0.068$)

Results: Shark Presence v Sea Surface Temperature





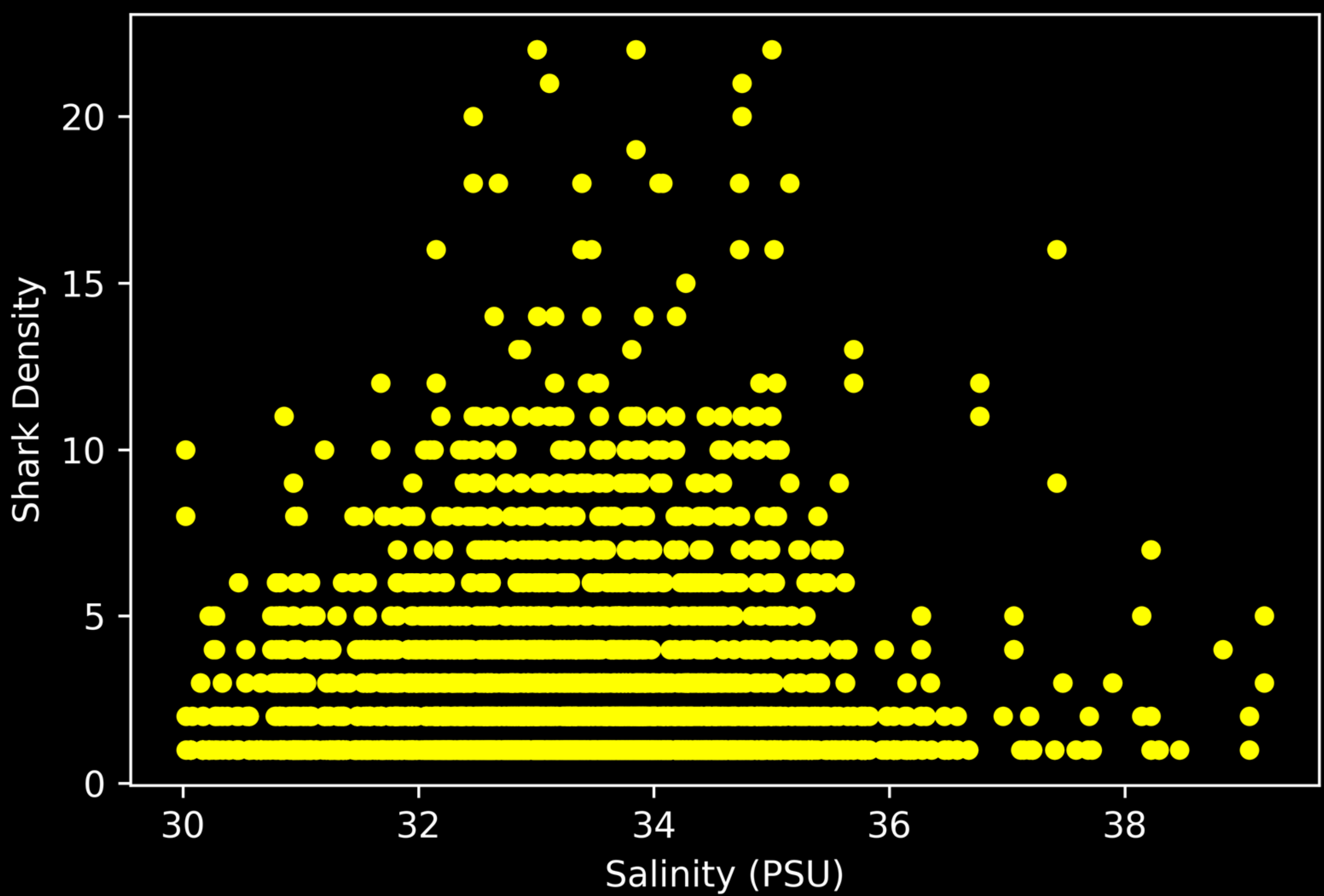


Trends in variance dissolve at the monthly scale

**Data in this analysis used method 1 for jackknifing*

Results: Shark Density v Sea Surface Salinity

**Data in this analysis used method 1 for jackknifing*

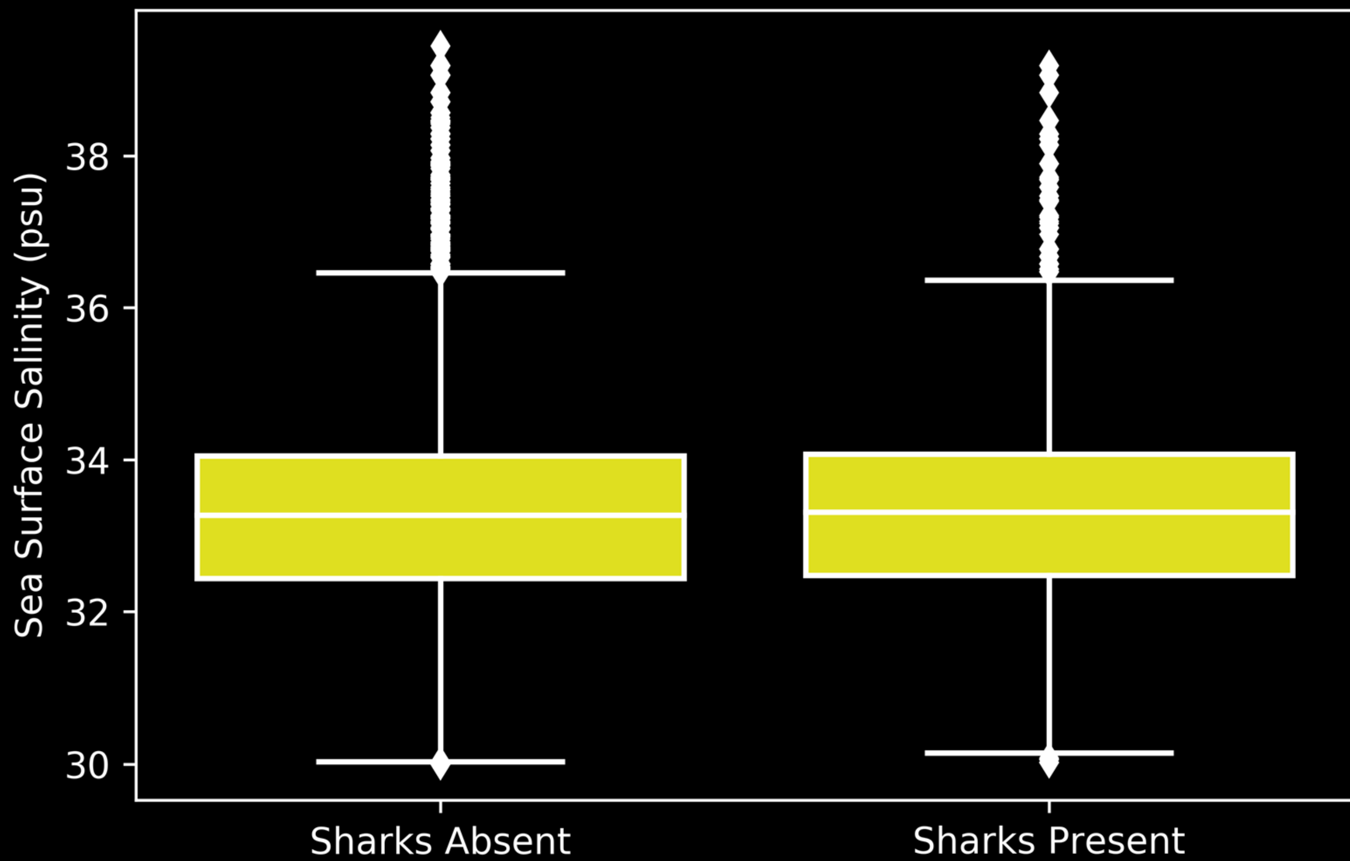


Min: 30.00
Max: 39.44
Median: 33.27
STD: 1.34

Highly variable shark densities at all salinity values

Results: Shark Presence v Sea Surface Salinity

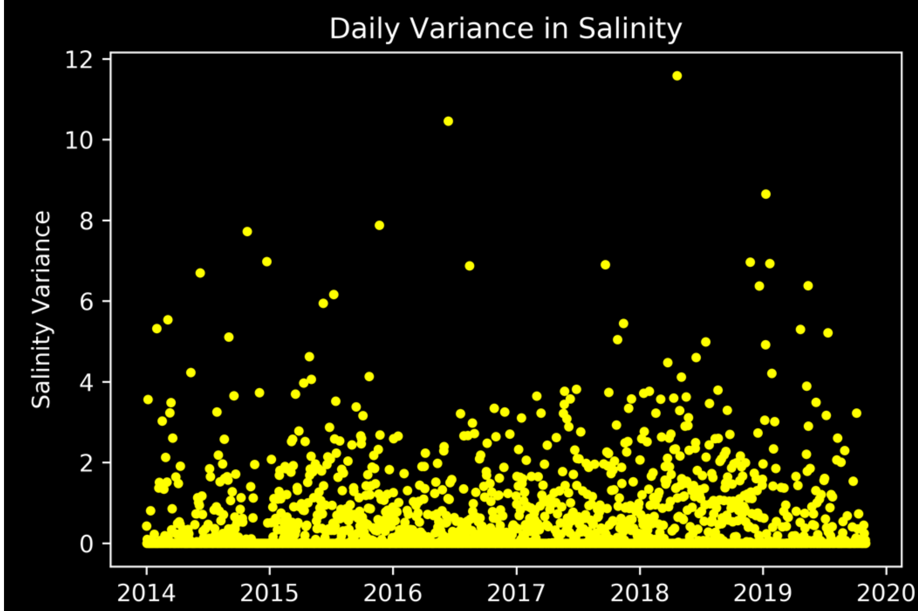
**Data in this analysis used method 1 for jackknifing*

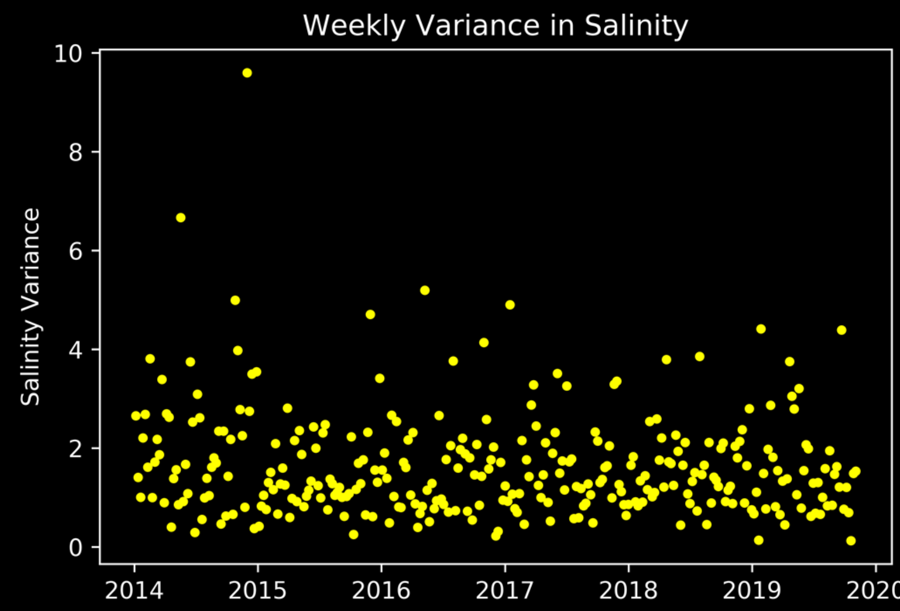


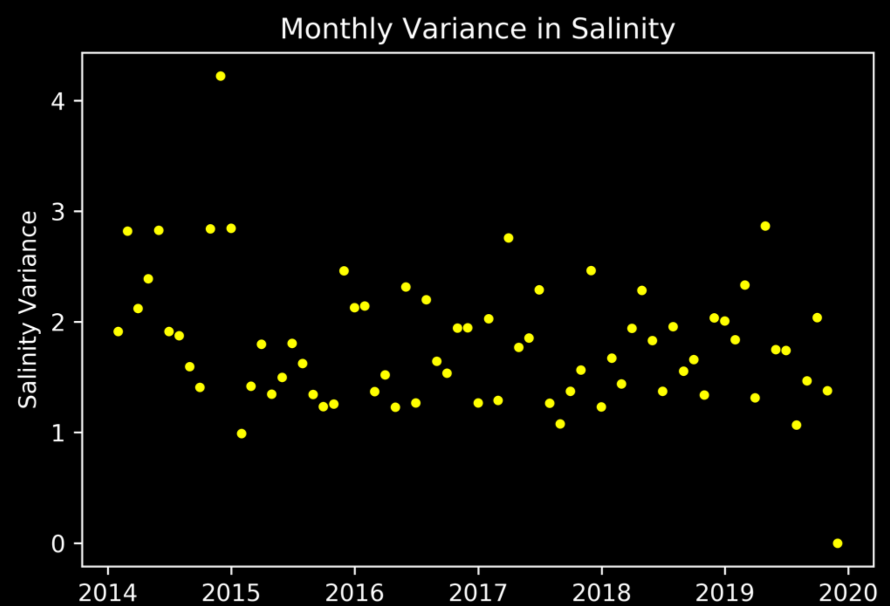
Mean (± SD): 33.29 ± 1.34 psu   33.29 ± 1.27 psu

No significant difference ($t = 0.199$, $p = 0.842$)

Results: Shark Presence v Sea Surface Salinity





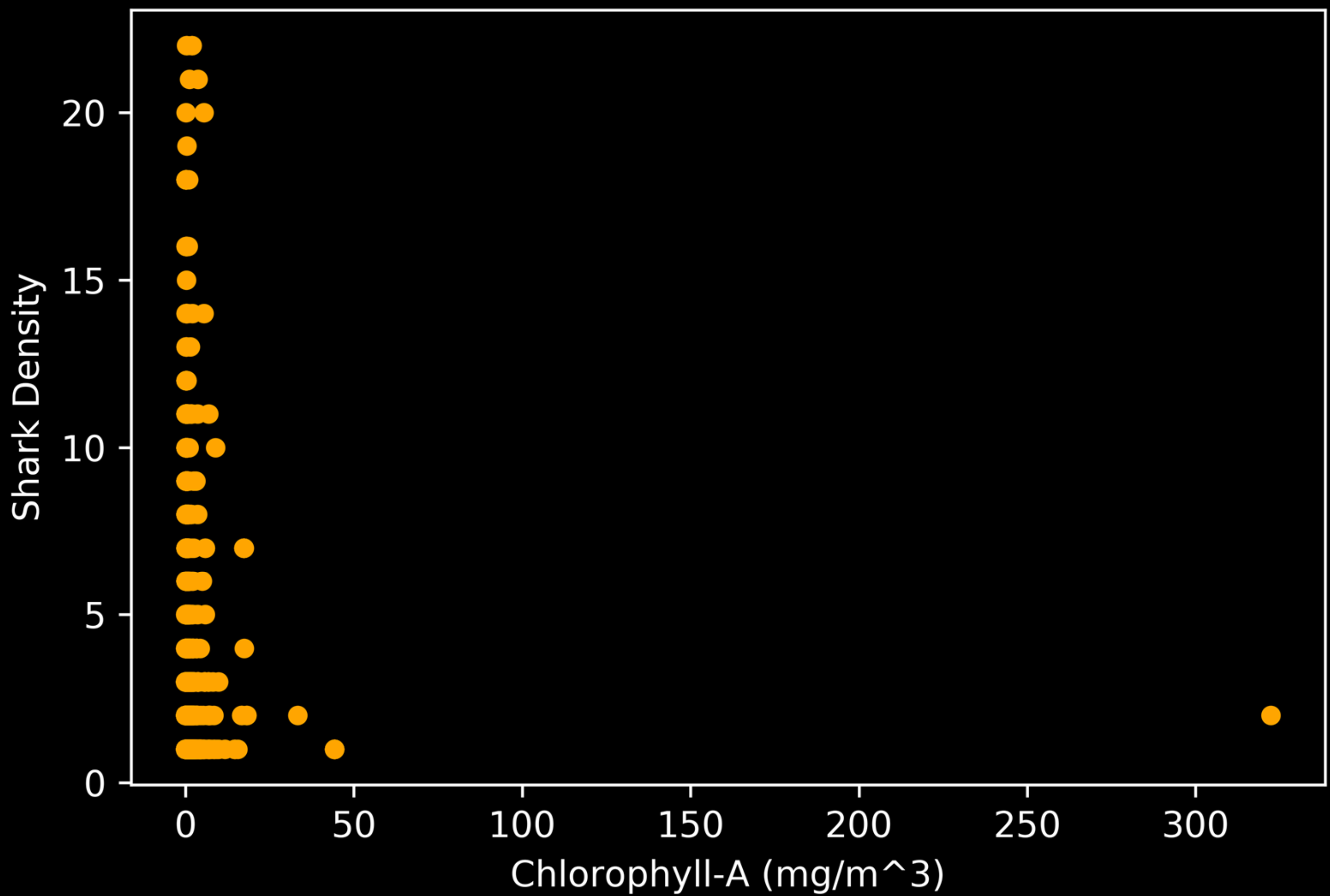


No clear trends in variance at any time scale

**Data in this analysis used method 1 for jackknifing*

Results: Shark Density v Chlorophyll-A

**Data in this analysis used method 1 for jackknifing*

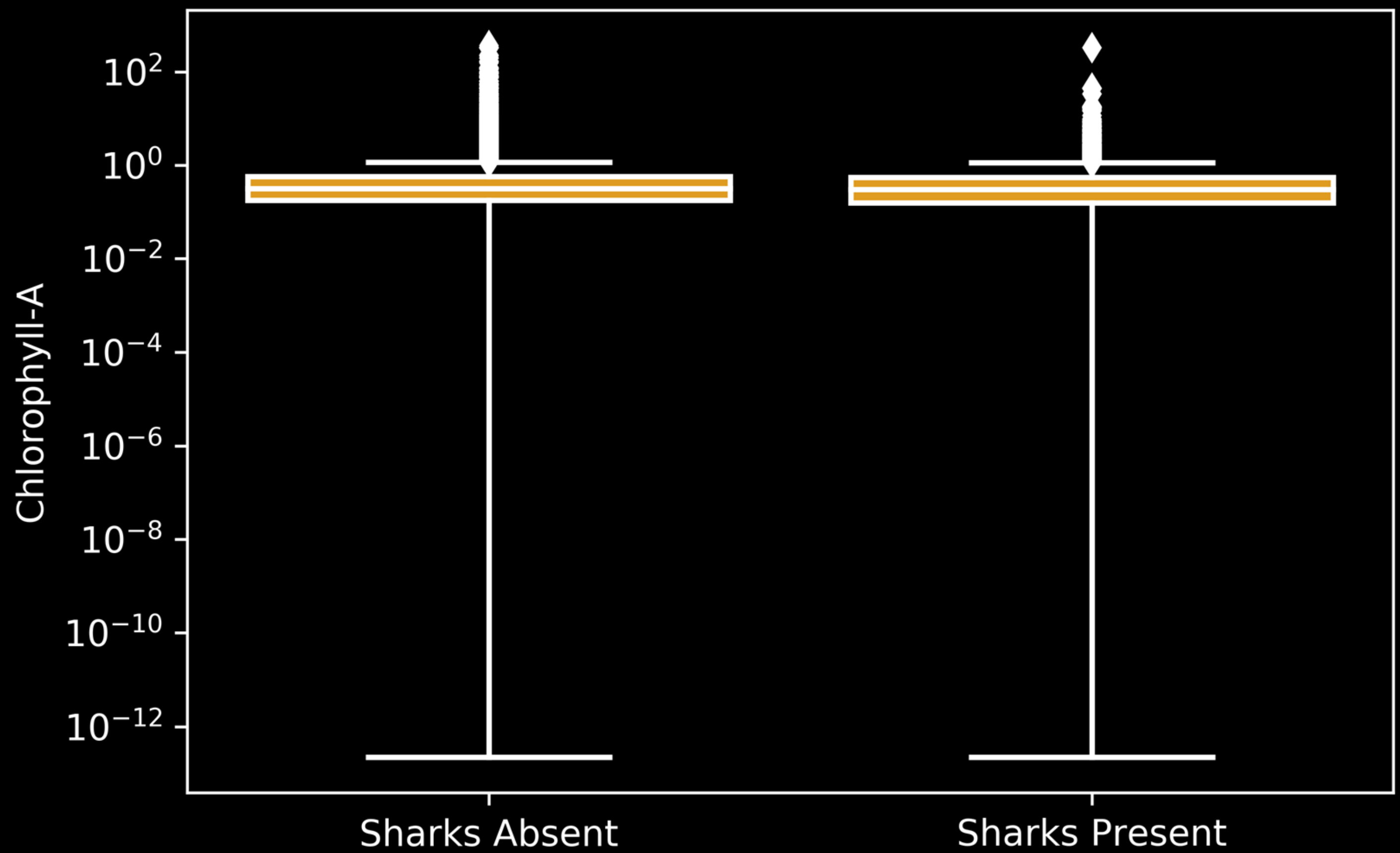


Min: 2.27e-14
Max: 359.25
Median: 0.32
STD: 7.38

Highly variable shark densities at chlorophyll values < 50

Results: Shark Presence v Chlorophyll-A

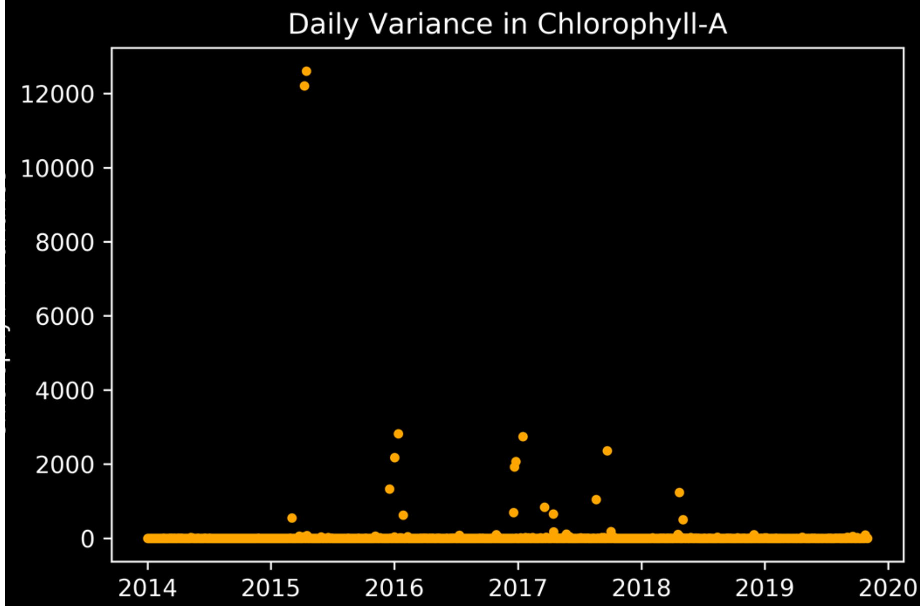
**Data in this analysis used method 1 for jackknifing*

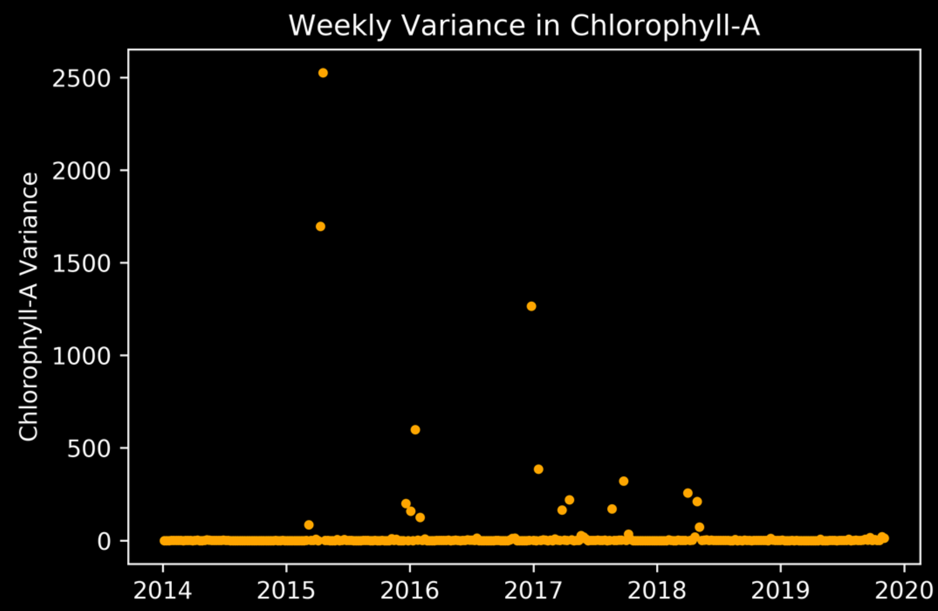


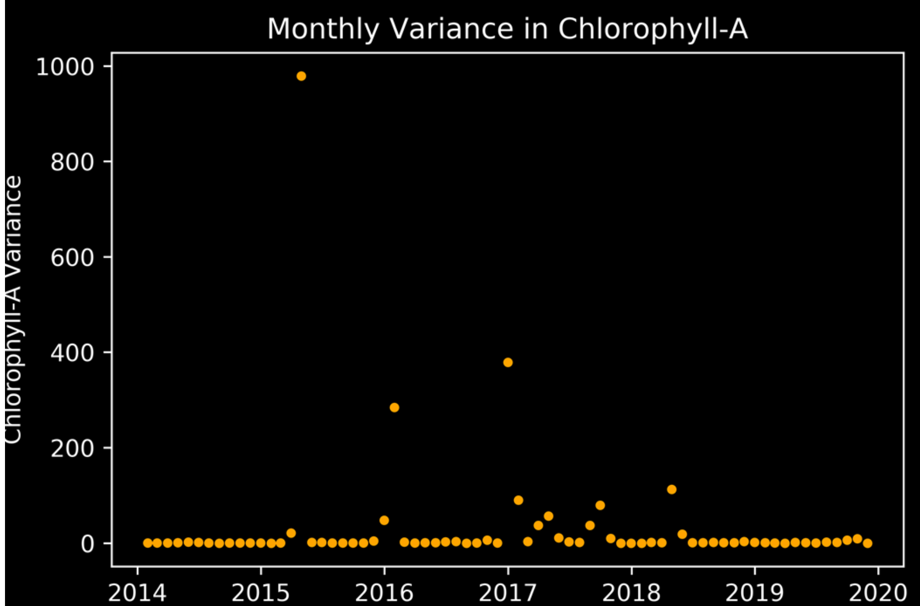
Mean (± SD): 0.86 ± 7.50 mg/m^3 0.72 ± 5.46 mg/m^3

No significant difference ($t = -1.119$, $p = 0.263$)

Results: Shark Presence v Chlorophyll-A





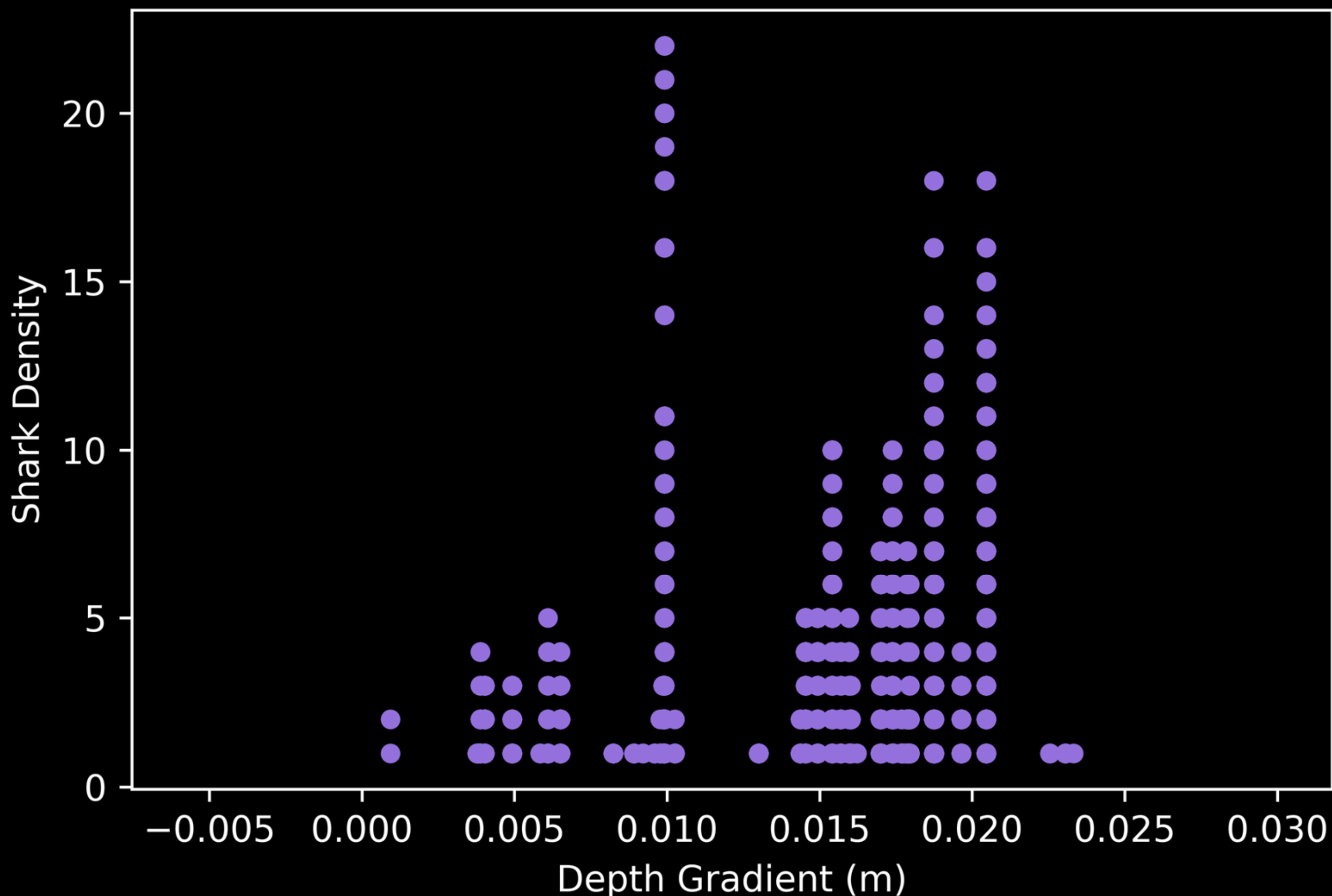


No clear trends in variance at any time scale

**Data in this analysis used method 1 for jackknifing*

Results: Shark Density v Seafloor Depth Gradient

**Data in this analysis used method 1 for jackknifing*

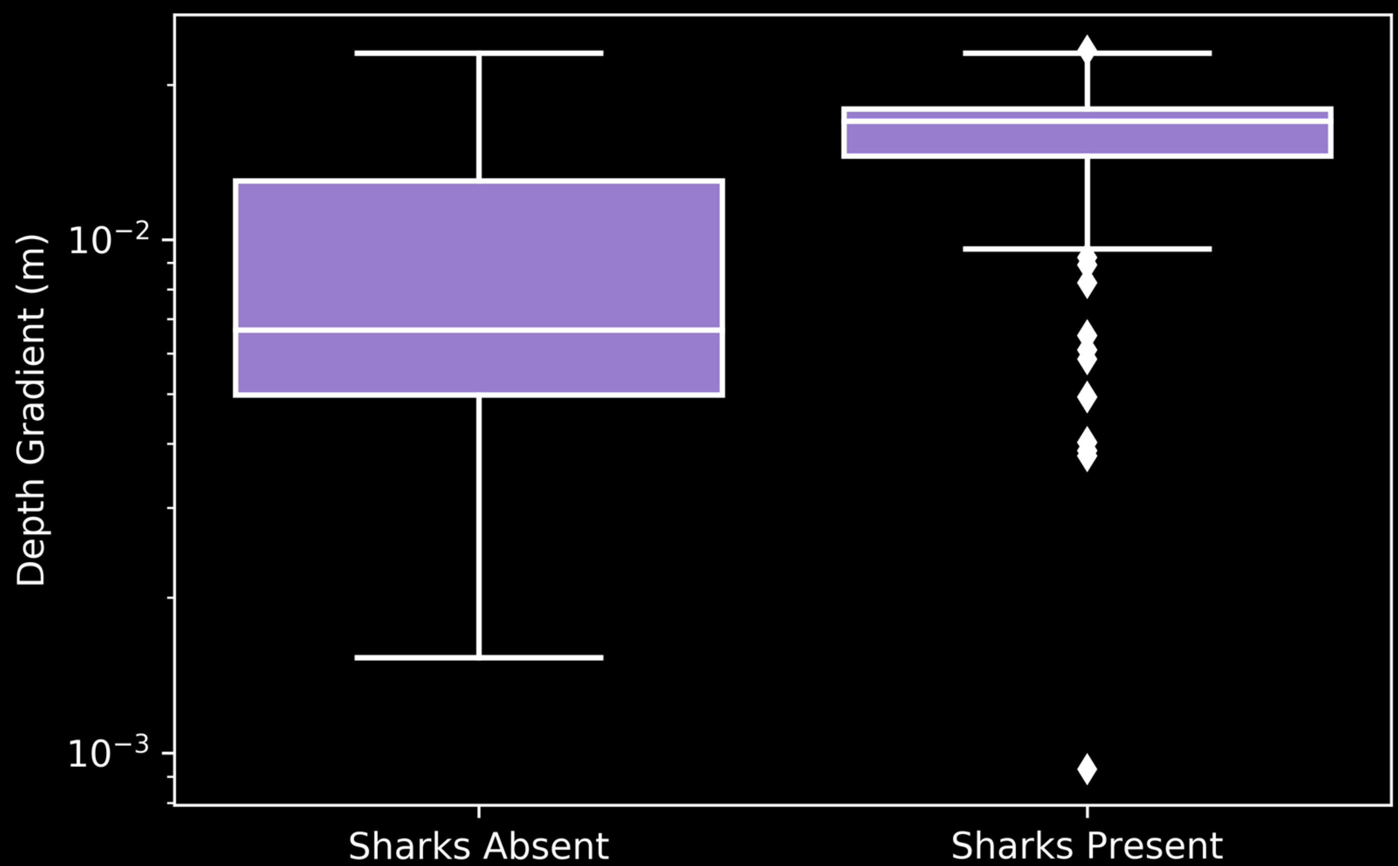


Min: 0.0009
Max: 0.023
Median: 0.007
STD: 0.005

Highly variable shark densities most gradient values

Results: Shark Presence v Depth Gradient

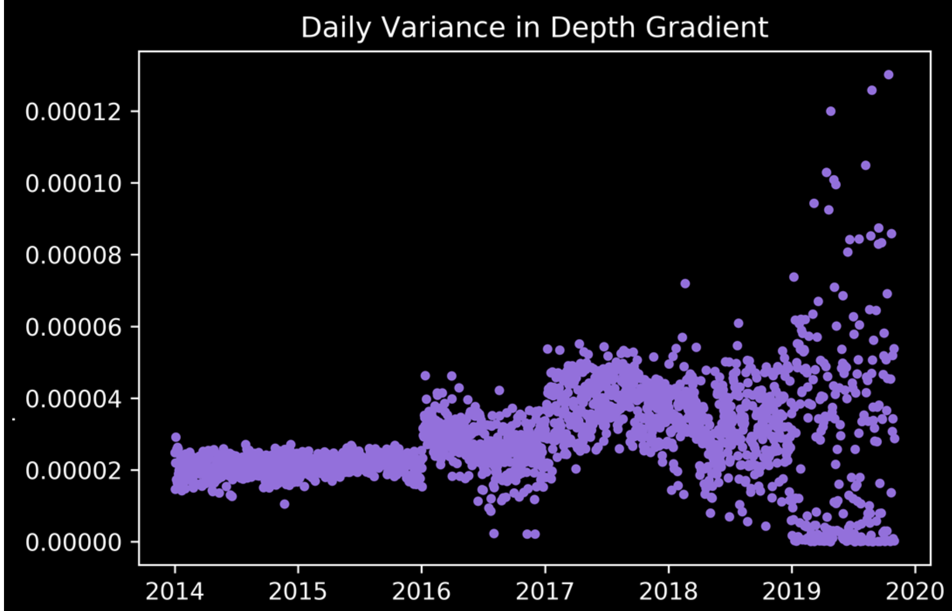
**Data in this analysis used method 1 for jackknifing*

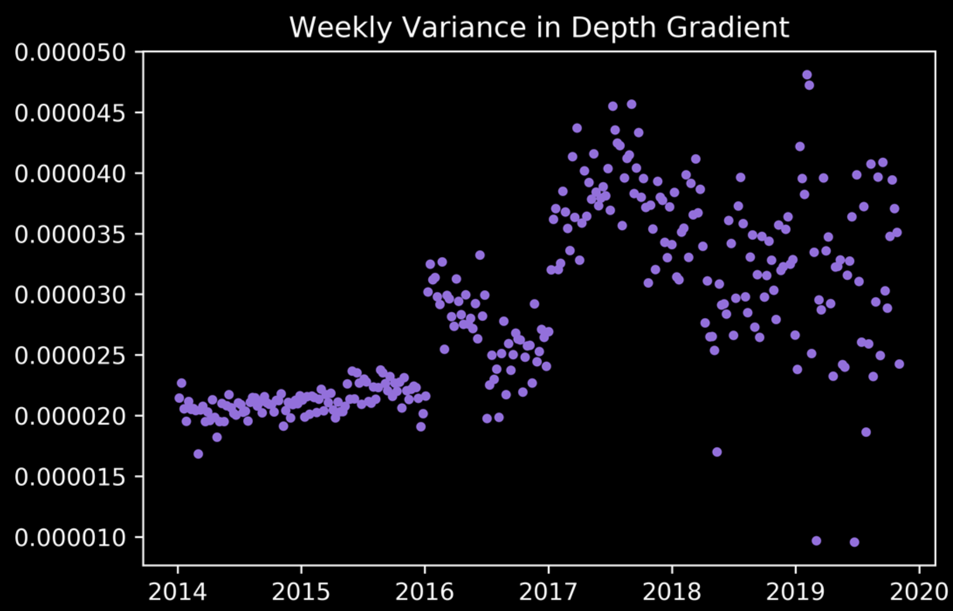


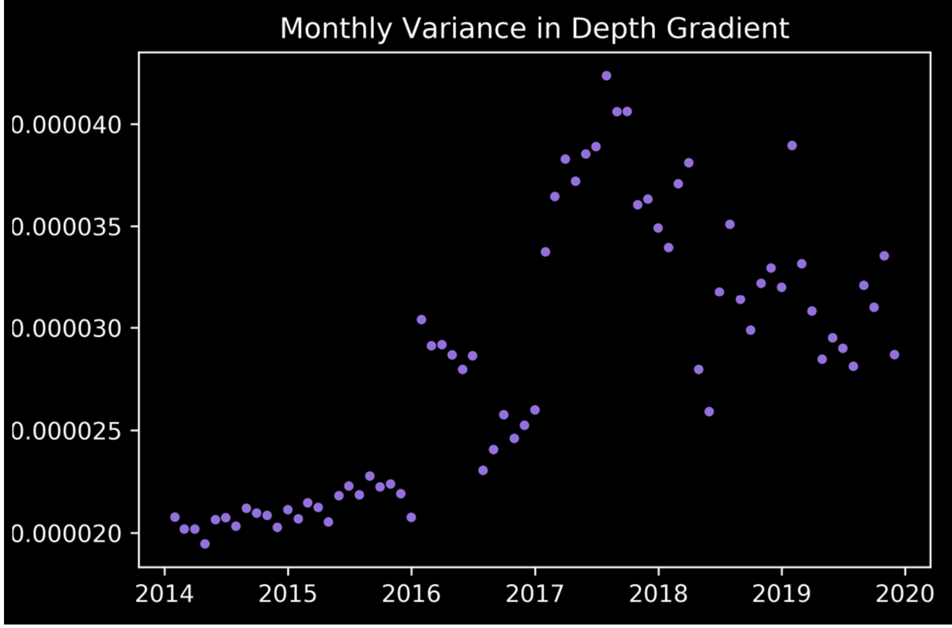
Mean (± SD): 0.009 ± 0.005 m   0.015 ± 0.004 m

Significant difference (t = 84.668, p < 0.01)

Results: Shark Presence v Depth Gradient







Increasing variance in recent years on all scales

→ More receiver coverage in steeper regions

**Data in this analysis used method 1 for jackknifing*

Methods: Preliminary Machine Learning

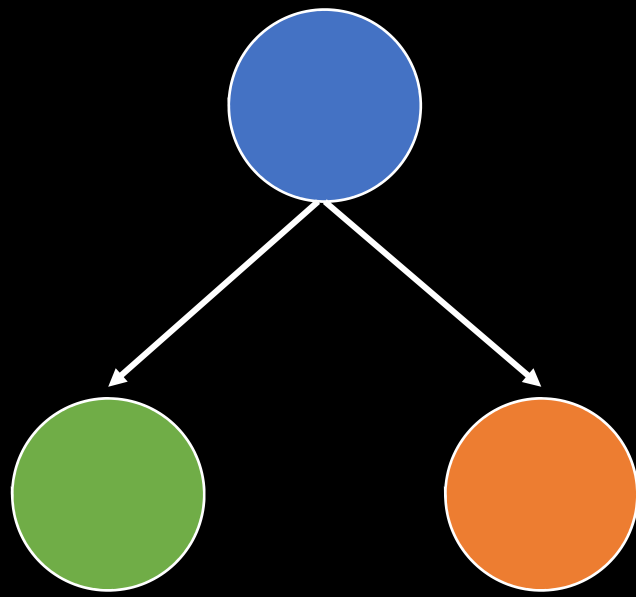
**Data in this analysis used method 1 for jackknifing*

Data Formatting:

- One Hot Encoding using *pandas* get_dummies()
  - Zone, Month, Year, Lunar Phase
- Data split into training and testing data

Preliminary Machine Learning:

- Decision Tree
- Gradient Boosting
- Stochastic Gradient Descent
- K Nearest Neighbors
- Gaussian Naïve Bayes
- Gaussian Process Classifier



Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

Decision Tree

Tested:

- Balanced vs non balanced class weights, range of max depths (grid search and *for* loop)

Best model (based on accuracy):

- Class weights = None, max_depth = 10
- Precision = 0.996; Recall = 0.985

Good performance, but hard to interpret after 10 levels.

Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

Gradient Boosting

Tested:

- Range of learning rates, max depth, and n estimators (*for* loop and grid search)

Results:

- Learning rate < 0.5 → higher accuracy
- Learning rate > 0.5 → lower accuracy

Best model (based on accuracy):

- Learning_rate = 0.075, max_depth = 3, n_estimators = 100

Harder to visualize decision splits.

Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

Stochastic Gradient Descent

Tested:

- Attempted grid search w/ several parameters

Results:

- Model was running for 12+ hours before termination
- Model ran with default hyperparameters
  - Precision = 0.99, Recall = 0.99

Unable to establish hyperparameters but yielded good results

Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

K Nearest Neighbors

Tested:

- Nearest Neighbors ranging from 1-30, uniform and distance weighting (*for* loop)

Results:

- Uniform Weighting with 10 nearest neighbors
  - Testing Precision = 0.991, Recall = 0.991
- Distance Weighting with 9 nearest neighbors
  - Testing Precision = 0.9925, Recall = 0.9945

Unable to describe classification in terms of predictors

Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

Gaussian Naïve Bayes

Tested:

- Default hyperparameters

Results:

- Testing Precision = 0.9999, Recall = 0.9337

Unable to establish hyperparameters but yielded good results

Results: Preliminary Machine Learning

**Data in this analysis used method 1 for jackknifing*

Gaussian Process

Tested:

- Default hyperparameters

Results:

- Too large for classifier to handle
- Fitted data with subsamples (5000 rows) at a time
- Testing Precision = 0.95, Recall = 0.998

Unable to establish hyperparameters but yielded good results

Methods: Refined Machine Learning

*Based on the preliminary machine learning analyses, these results seemed **too good to be true.***

***Hypothesis:** Models are automatically classifying regions with receiver densities of 0 to have no sharks*

*Technically true, but **high potential of false negatives.***

***Response:** Re-do jackknifing to only include times when receiver density > 0.*

Methods: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Data Formatting:

- Jackknife data to only include receiver densities > 0
- Manually balance data:
  - Additional random samples of data when sharks are present (not enough data to down-scale)
- One Hot Encoding using *pandas* get_dummies()
  - Zone, Month, Year, Lunar Phase
- Data split into training and testing data

Machine Learning Algorithms

Decision Tree

Gradient Boosting

K Nearest Neighbors

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Decision Tree

Tested:

- Max depth between 3 and 10 (grid search)

Best model (based on precision):

- Max depth = 10

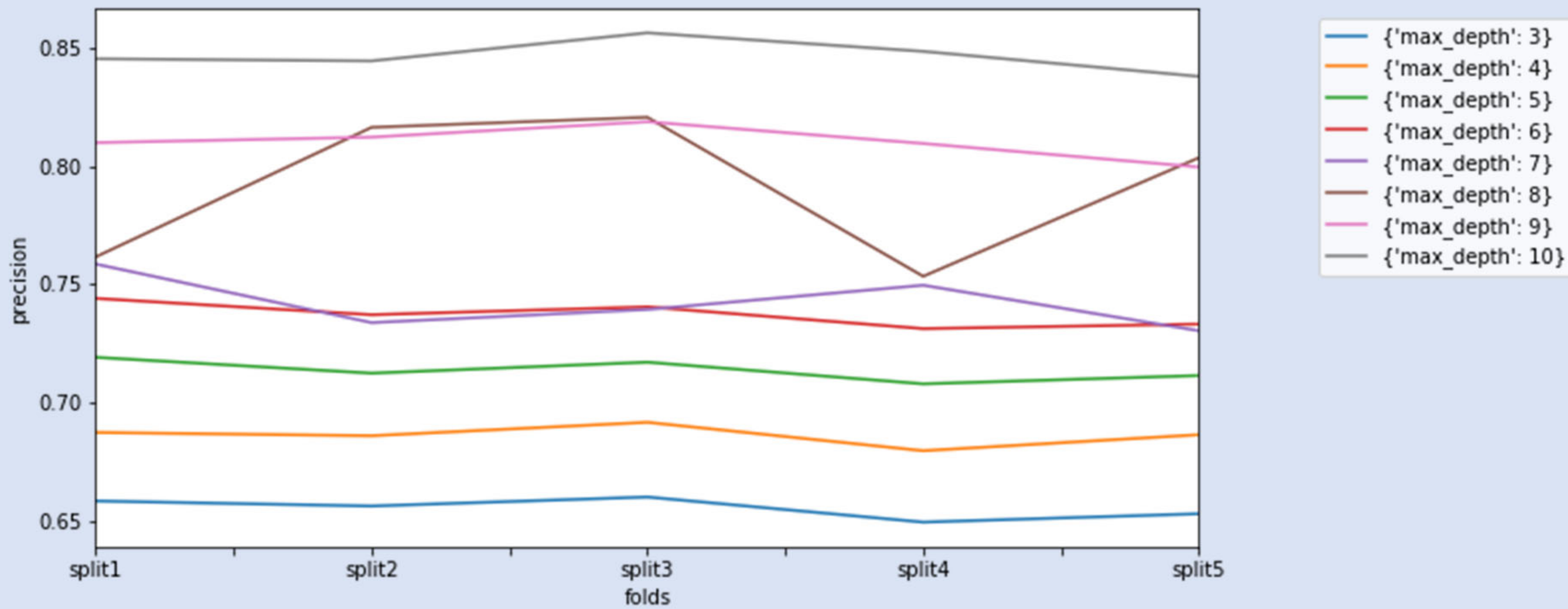
But all models showed precision between 0.65 and 0.85

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Decision Tree

All models showed precision between 0.65 - 0.85



Max_depth = 6 → Easier to interpret

Training data precision = 0.928

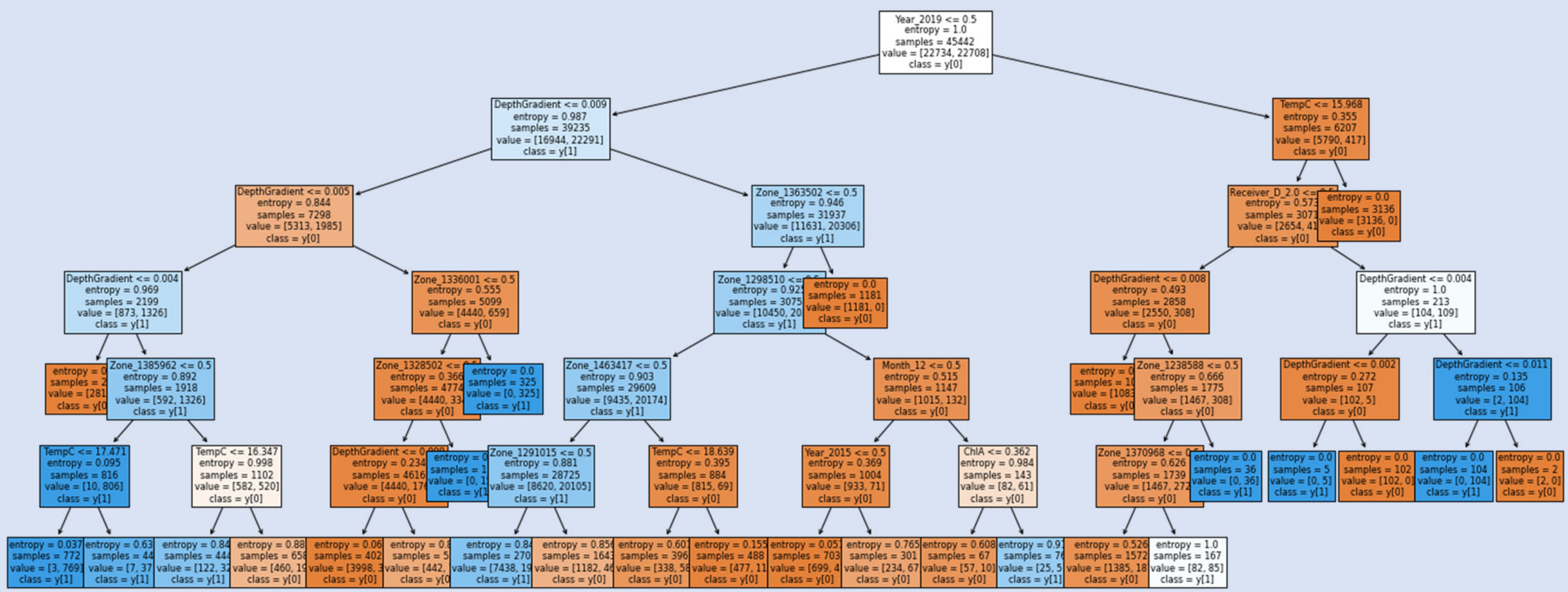
Testing data precision = 0.931

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Decision Tree

Max_depth = 6 → Easier to interpret



Most important parameters:

Depth gradient; year 2019; receiver density = 2, temperature, particular zones

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Gradient Boosting

Tested:

- Learning rate, max depth, n estimators (grid search)

Best model (based on precision):

- Learning_rate = 0.75, max_depth = 8, n_estimators = 100

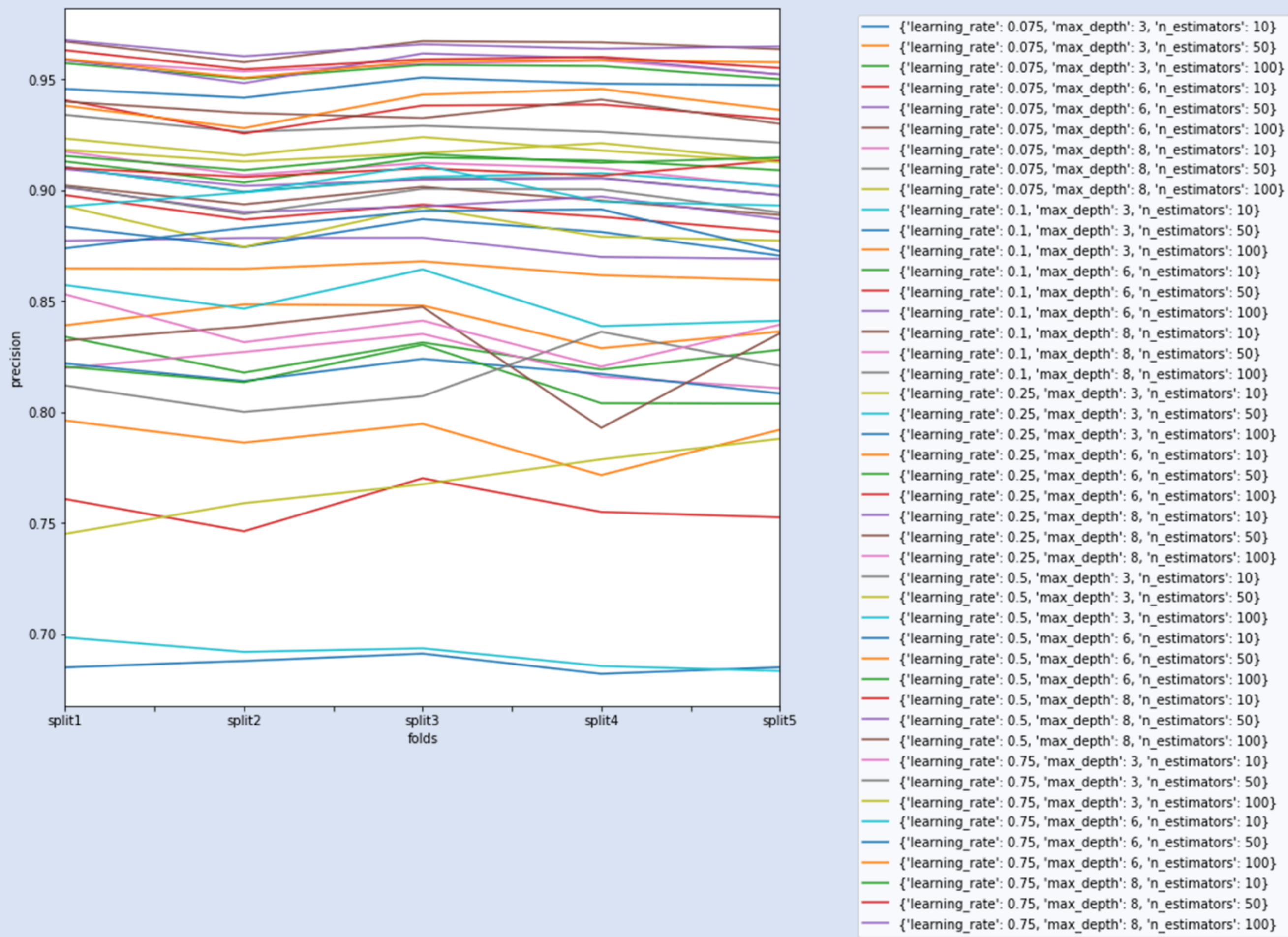
But all models showed precision between 0.65 and 0.99

Results: Refined Machine Learning

*Data in this analysis used method 2 for jackknifing

Gradient Boosting

All models showed precision between 0.65 - 0.99



Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Gradient Boosting

Simplest model with decent precision:

Learning_rate = 0.075, max_depth = 3, n_estimators = 100

Training data precision = 0.904

Testing data precision = 0.903

Most important parameters:

Depth gradient; year 2019, 2015, 2018; receiver density = 2; temperature; particular zones

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Gradient Boosting

How does this model perform without receiver density as a predictor?

- Estimate shark presence when receivers aren't present

Best model (based on precision):

- Learning_rate = 0.75, max_depth = 8, n_estimators = 100

But all models still showed precision between 0.65 and 0.99

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

Gradient Boosting (without Receiver Density)

Simplest model with decent precision:

Learning_rate = 0.075, max_depth = 3, n_estimators = 100

Training data precision = 0.916

Testing data precision = 0.9156

Most important parameters:

Depth gradient; year 2019, 2015, 2018; temperature; particular zones

Results: Refined Machine Learning

**Data in this analysis used method 2 for jackknifing*

K Nearest Neighbors (without Receiver Density)

Tested:

- Nearest Neighbors ranging from 1-30, uniform and distance weighting (grid search)

Best model (based on precision):

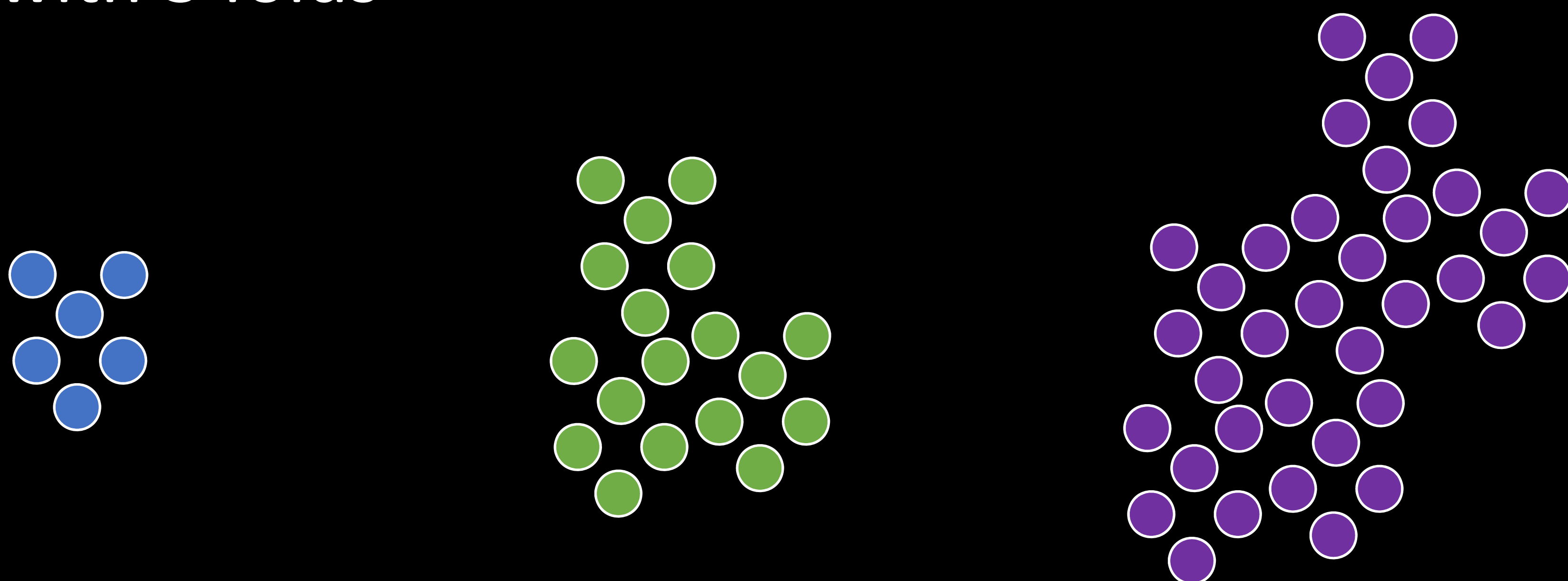
- Nearest neighbors = 1, uniform weighting

Testing data precision = 0.999

Methods: Optimal Sample Size

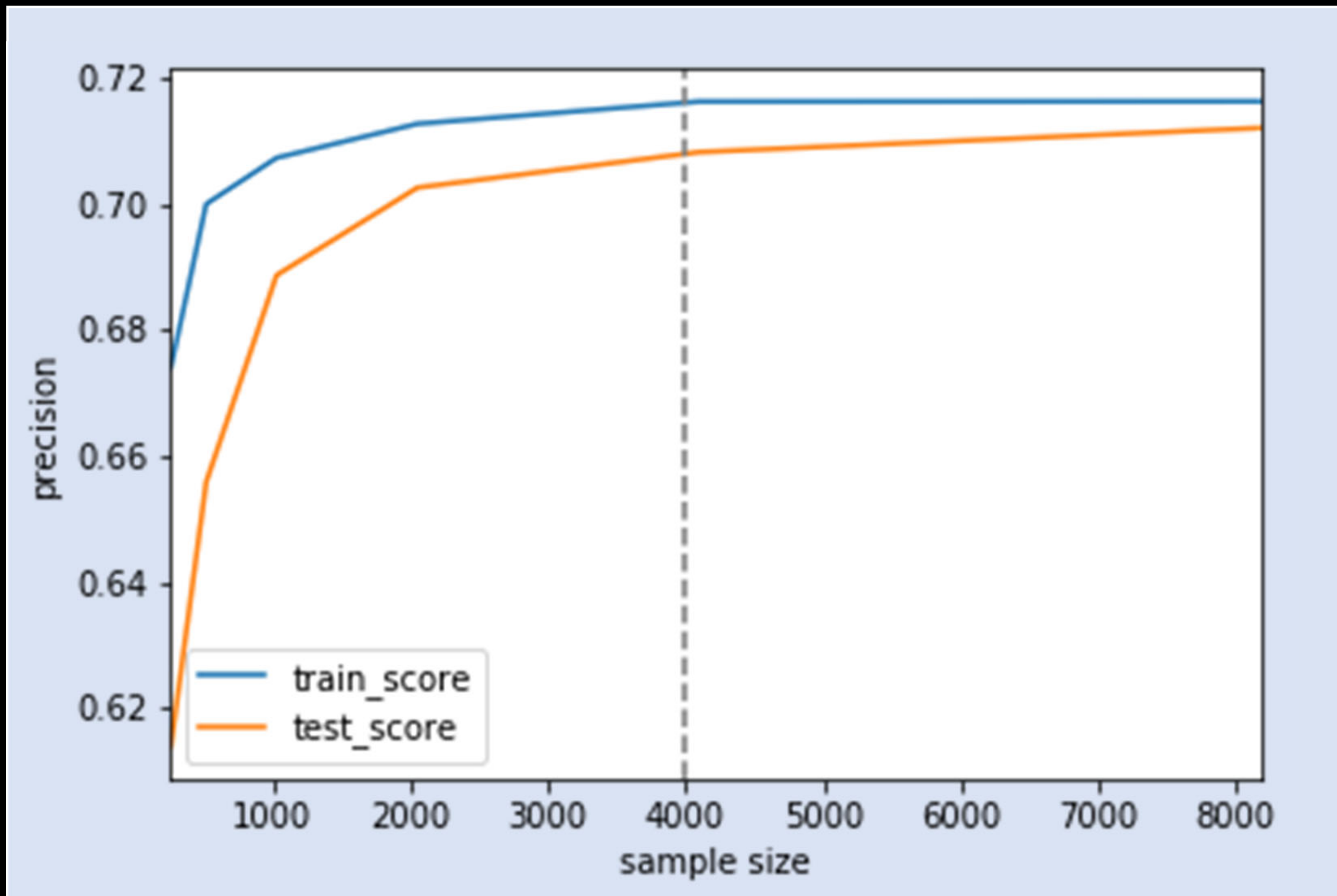
**Data in this analysis used method 2 for jackknifing*

- Sample sizes using 2^8 … 2^n … 2^{14} random samples
- 100 iterations for each sample size
- Tested using the Decision Tree Classifier
- Average precision score using *cross_val_score* with 3 folds



Results: Optimal Sample Size

**Data in this analysis used method 2 for jackknifing*



Precision scores plateau around 4000 samples

Discussion: Choosing the Best Model

- Best model should reduce the probability of false negatives (training data **only** where receivers are present)

- ***Likely depends on the audience***
  - Users that want to be able to understand how the decision was made
    - Decision Tree: Lay out the exact results but with lower confidence

  - Users that want just “shark likely present” vs “shark likely absent” with precision estimate
    - Gradient Boosting (high confidence)
    - K Nearest Neighbors (highest confidence, longer testing time)

Discussion: Choosing the Best Model

- ***Likely depends on the audience***
  - Users that want to figure out where to put receivers
    - Gradient Boosting or K Nearest Neighbors using previous environmental data for potential regions
  - Users that want to fill in gaps in historical data
    - K Nearest Neighbors (highest confidence)
  - Users that are studying a different species
    - Decision Tree can predict sound results at smaller sample sizes (n = 4000)

Discussion: Finding the Best Parameters

- Shark presence at steeper depth gradients (T-Test; Decision Tree Classifier)
  - Juvenile White Sharks feed on soft-bodied prey in sediment[1]
  - Might be easier to acquire prey at slightly steeper slopes

- Yearly variance, particularly 2019
  - 2019 was not yet over when project began
  - Each year has unique environmental characteristics
  - Makes filling in historical data easier
  - Makes predicting future data harder

1: Estrada, J. A., Rice, A. N., Natanson, L. J., & Skomal, G. B. (2006). Use of isotopic analysis of vertebrae in reconstructing ontogenetic feeding ecology in white sharks. Ecology, 87(4), 829-834.