Disease Outbreak Dictates Spatiotemporal Variation in Ochre Star Abundance on the US West Coast

Paige Duffin; MADA Term Project

2019-12-06

# Abstract

Outbreaks of infectious disease are increasingly recognized as key regulators of host populations across time and space. Sea Star Wasting Disease (SSWD) is an epizootic crisis facing modern day coastal ecosystems, as outbreaks of asteroid wasting trigger mass mortality events across a broad range of sea star species, including *Pisaster ochraceus*, a keystone predator of the rocky intertidal. This investigation makes use of publicly available data from MARINe, a monitoring network aimed at capturing changes in shoreline habitats along the west coast of the US. I utilized exploratory and modeling approaches to characterize changes in *P. ochraceus* populations in the years before, during, and after the recent and devastating SSWD outbreak of 2013-2014. I observed trends that suggested Ochre sea star populations suffered high rates of loss during the recent disease event, but may be capable of recovery through recruitment of young individuals to select locations.

# Introduction

## Sea Star Wasting Disease: A Modern Epidemic

Outbreaks of infectious disease are an important, yet poorly understood, driving force in population biology and, while prevalent in both terrestrial and marine systems, the manner by which outbreaks influence biotic relationships, age demographics, community structure and function, and trophic interactions in the aquatic realm consistently lags behind that of terrestrial disease ecology. The growing body of evidence in this field, however, supports the observation that marine disease epidemics are increasing in frequency and severity (Orth et al., 2006) (Waycott et al., 2009). Elucidating the relationship between aquatic species and their pathogenic diseases is pertinent in the field of marine ecology because infection outbreaks have the potential to drastically alter ecosystem functionality (Burge et al., 2014). Many of the organisms which have faced chronic and/or severe outbreaks of disease (such as sea urchins, scleractinian corals and seagrasses) are also considered keystone species and/or ecosystem engineers, meaning that they contribute highly valuable services or functions to their surrounding community and ecosystem. Thus, it comes as no surprise that disease-driven mass mortalities of these species often generate waves of ecological permutation, ranging from temporary local disruptions to permanent phase shifts (Burge et al., 2014), such as the transformation from a coral to macroalgal dominated Caribbean reef structure that accompanied the massive *Diadema antillarum* (black sea urchin) die-off of the early 1980s (Lessios, 1988). This event, likely caused by a biological pathogen (Schultz, Cloutier, & Côté, 2016), reduced Diadema populations to less than 1% of their original size (Lessios, 1988). Mass mortality events impacting critical foundation species, ecosystem engineers, or keystone species such as the black sea urchin have been coined ‘marine disease emergencies’ due to the detrimental cascade of events that often succeeds them (Miner et al., 2018).

Sea Star Wasting Disease (SSWD) is another epizootic crisis facing modern day coastal ecosystems. Outbreaks of asteroid wasting have been documented periodically since the late 1970s and SSWD describes a suite of symptoms observed across a broad range of sea star species, most of which play integral parts in shaping their community structure (Eisenlord et al., 2016) (Bucci et al., 2017) (Miner et al., 2018). Generally speaking, the wasting disease events which punctuated the past four decades were relatively brief, in localized areas, and largely failed to capture the attention of the scientific community. Beginning in summer 2013, however, mass mortalities of sea stars due to wasting disease have caused unprecedented damage, owing to the geographical and temporal extent of impact (Eisenlord et al., 2016) (Bucci et al., 2017) (Miner et al., 2018). This epizootic event, which has killed millions of asteroids across over 20 taxa, is widely referred to as the largest disease event sweeping through a wildlife marine species in documented history (Hewson et al., 2014) (Gudenkauf & Hewson, 2015) (Eisenlord et al., 2016). There is some discrepancy in the literature as to when the modern epidemic ended, with a handful of literature citing the duration as a one-year span, from 2013 to 2014 (Kohl, McClure, & Miner, 2016)(Montecino-Latorre et al., 2016), while others assert it occurred until 2015 (e.g. (Bucci et al., 2017)); a final set of articles document the outbreak as ongoing or remain vague about specific time frames (e.g. (Eisenlord et al., 2016)). For the purpose of this investigation, I will be operating under the assumption that the timeframe of SSWD spanned from mid-to-late 2013 to early 2014, and will refer to this event as the “SSWD event” or the “disease event” hereafter.

The consensus of reports which emerged in the years after the disease event indicates that the ochre sea star, *Pisaster ochraceus*, suffered very high mortality (Eisenlord et al., 2016)(Schiebelhut, Puritz, & Dawson, 2018)(Miner et al., 2018). Ochre stars occupy the lower to mid-intertidal on rocky shores, spanning an impressive range from Alaska to Baja California (Harley, Pankey, Wares, Grosberg, & Wonham, 2006), and as keystone predators, the health of *P. ochraceus* populations heavily influence the surrounding community structure and species richness (Harley et al., 2006). There has been a special emphasis on the characterization of SSWD in this species, and researchers are highly motivated to address questions like: what can we learn from the individuals who survived and/or from the populations which have rebounded? How has community structure been altered as a result of this epidemic? Are the results at different locations similar? If not, what drives the differences? Though devastating, disease outbreak events such as this provide unique opportunities in the coming years to better understand how a suite of ecologically important organisms respond to unprecedented levels of ecological permutation.

## Description of data and data source

### *The Source*

The vast majority of sea star abundance surveys and wasting disease documentation cover the west coast of the United States. While it is almost certainly true that SSWD is most prevalent in this region, there is a strong bias towards the discovery of diseased organisms owing to the emphasis of long-term survey networks in the area, such as the extensive [Multi-Agency Rocky Intertidal Network](https://marine.ucsc.edu/index.html) (MARINe for short). MARINe is a consortium of organizations (lead group: University of California Santa Cruz) which began with the goal of developing a set of methods that allows scientists to collect data that is statistically sound with approaches that are simple, cost-effective, reproducible, and sustainable. Among their initiatives and projects is the targeted long-term monitoring approach which focuses on monitoring key species within fixed plots semi-annually, with the broad aim of collecting data which allows for the detection of relatively small changes in the abundance of species which play an important role in their community. Specifically, the data set utilized in this investigation focuses on the long-term abundance of three target species: ochre sea stars (*Pisaster ochraceus*), the mottled sea stars (*Evasterias troschelii*), and the black chitons (*Katharina tunicata*).

The data used in this analysis is publicly available by request through the UC Santa Cruz [MARINe portal](https://marine.ucsc.edu/explore-the-data/contact/index.html).

The purpose of this investigation is to make use of this extensive data monitoring system to investigate spatial and temoporal changes in target species assemblages along the west coast of the United States, with particular emphasis on *P. ochraceus* populations in the years before, during, and after the SSWD outbreak of 2013-2014.

### *The Data*

After the data request was approved, the raw data was sent by email and downloaded as a .csv file (seastarkat\_size\_count\_totals.csv). Following the cleaning step (see methods), the finalized data was comprised of 12165 observations across 23 variables. This included the following (field definitions obtained from accompanying MARINe codebook file):

* 14 categorical variables (factors)
  + **georegion:** geographic region in which site is located
  + **bioregion:** biogeographic region in which site is located
  + **island:** the name of the island where the referenced site is located. Sites not on islands are designated as mainland.
  + **state:** state where the referenced site is located.
  + **mpa\_region:** describes whether the referenced site is located within a Marine Protected Area (MPA).
  + **marine\_site\_name:** the name of the site where the survey was conducted.
  + **site\_code:** abbreviated name of the site where the survey was conducted.
  + **season\_name:** name of season when the survey was conducted.
  + **marine\_season\_code:** four-character code to identify the Sampling season. The first two characters indicate the season and the last 2 characters indicate the year.
  + **species\_code:** unique alphabetic lookup code for each species or lumped species group.
  + **group\_code:** the unique code for each monitoring group
  + **group\_code\_UCSC\_other:** modified groupings of levels in group\_code, see processing\_script for details.
  + **method\_code:** the method used for sampling the plot.
  + **method\_code\_IP\_other:** modified groupings of levels in method\_code, see processing\_script for details.
* 9 numeric variables (integers/numeric)
  + **latitude:** the number of degrees North of the equator, expressed in decimal degrees to 5 decimal places (NAD83).
  + **longitude:** the number of degrees West of the Prime Meridian, expressed in decimal degrees to 5 decimal places (NAD83).
  + **marine\_sort\_order:** used to order the sites geographically along the coast. Actual values are arbitrary, and should not be used for site identification.
  + **marine\_common\_year:** year in which survey was done. Note that this can be different from the year in the survey date - for example, January of 2016 could still be considered FA15, so the marine\_common\_year would be 2015.
  + **marine\_common\_season:** consecutive sampling number. A code created to allow sequencing of the seasons, used for sorting data in chronological order.
  + **season\_sequence:** season sequence within each year. Each number indicates a different season.
  + **total:** total number of individuals counted in a given size\_bin.
  + **size\_bin:** size of the species being counted, binned to the nearest 5 or 10 millimeter.
  + **size\_sort\_order:** numerical code used for sorting size\_bin.

## Questions to be addressed

### *Exploratory Questions*

**1. How does species abundance change across time and space?**

**2. Are there discernable changes in spp. abundance before & after the focal SSWD event?**

**3. How did populations of the focal sp., *Pisaster ochraceus*, respond & recover from the SSWD event?**

### *Statistical/Machine Learning Questions*

**1. Which variables are good predictors of species abundance?**

**2. Are there definable geographical regions based on the available predictor variables?**

# Methods and Results

***Note:*** *due to the abundance of figures included in this report (over 30), I placed the figures in a second file (Duffin-Project-Manuscript-Figures.Rmd, which knits to a word document) so as not to disrupt the body of the text. Figures described throughout this text should be referenced therein while reading this document.*

## Data Analysis Methods

### *Data Import and Cleaning*

The data used in this investigation was downloaded as a .csv file (seastarkat\_size\_count\_totals.csv), imported into RStudio, and cleaned extensively prior to use in this analysis. The code which performed this cleaning can be found here: ./code/processing\_code/processing\_script. Generally, this involved examining each variable using the functions skim() and glimpse(), then plotting factors as bar plots and integers/numerics as histograms. This allowed me to observe the structure of each variable and make necessary changes to clean the data.

### *Analysis and Visualization Tools*

All analyses were conducted in R Studio (Version 1.2.1335) using various packages (details in indvidual scripts under code folder of project repository), though several were relied on heavily. Specifically, the following packages were used extensively throughout: caret, dplyr, ggplot2, ggridges, grid, gridExtra, knitr, skimr, tibble, tidyr, and tidyverse.

## Exploratory Analyses

The code which performed my exploratory analyses can be found here: ./code/analysis\_code/Exploratory\_Analysis.Rmd.

### *Mapping Survey Sites*

Geographical maps were generated to visualize the sample sites, which were distributed across four states along the west coast of the United States: Alaska (fig 1), Washington (fig 2), Oregon (fig 3), and California (fig 4). Overall, Alaska was surveyed the least-extensively (only one sample site; fig 1), and California -by far- the most, with sample sites lining the Pacific coast almost continuously (fig 4).

### *Question 1: How does species abundance change across time and space?*

#### Broad changes in species abundance

First, species abundance was investigated as a collective unit across all three species and over all years surveyed (2000-2018). Throughout this analysis, species abundance was plotted as log-transformed values to reduce the visual burden of outliers. Figure 5 displays trends over time as a function of state, clearly revealing that Alaska and Washington were not sampled before 2009. At this scale, it is hard to resolve any meaningful trends in species abundance across time and space (fig 5).

#### Abundance trends by species

As this investigation is primarily interested in one of the three species surveyed, *P. ochraceus*, it was logical to separate the data by species surveyed. Figures 6, 7, and 8 plot the log-transformed abundances of *P. ochraceus*, *K. tunicata*, and *E. troschelii* across all years in which the data existed. For P.ochraceus, this spanned from 2000 to 2018 (fig 6); for the other two species, sampling began in 2009 (figs 7-8). It became immediately clear that *E. troschelii* was relatively sparse as compared to *P. ochraceus* and *K. tunicata* (figs 6-8). *P. ochraceus* abundance appears to be relatively stable from 2000 to 2013, after which the median abundance notably drops (fig 6); this is exactly what we would expect to see given our prior knowledge, and is worth investigating further.

### *Question 2: Are there discernable changes in spp. abundance before & after the focal SSWD event?*

In an attempt to better visualize changes in spp. abundance before & after the focal SSWD event, the data was filtered to only include sampling events between 2013 and 2016. This subset of data was then plotted on a map to visualize the general ranges of each of the three target species (fig 9). While *P. ochraceus* dominates the visualization, we can see that *E. troschelii* is only present in the northern region of the continental US (Washington) and at the site in Alaska (fig 9). This narrower range of habitat likely contributes to the trends discussed in the previous section. In contrast, *P. ochraceus* appears to occupy the entire range of sites surveyed (fig 9).

Next, trends in abundances of each species were visualized between 2013 and 2016 using boxplots (figs 10-12). Again, the limited number of entries for *E. troschelii* prohibit any meaningful detection of changes in abundance (fig 12), but the narrowed timeframe allows us to see that median abundance values of black chiton (*K. tunicata*) populations are relatively uniform across the four years plotted (fig 11). Perhaps most interestingly, there is a pronounced dip in *P. ochraceus* abundance from 2013 to 2014, and even into 2015 (fig 10). Somewhat counterintuitively, however, outlier sites with high abundance seem to increase over time (fig 10). This will be investigated in more detail, below, under “Question 3.”

It is possible that averaging trends across a broad range of sampling locations might dilute real changes that are occuring, so in an attempt to tease apart the trends observed in figures 10 through 12, I subdivided each plot by state (fig 13a-c). Again, *E. troschelii* fails to produce any trends worth investigating (fig 13c), so it will not be included in downstream analyses. When separated by state, *K. tunicata* abundances appear to flucuate more than previously detectable, except in California (fig 13b). However, I speculate this may be due to relatively few observations remaining to plot for Alaska, Washington, and Oregon. Further, California again stands out among *P. ochraceus* plots in that it reflects the same general trend observed when averaged across states (fig 13a). *P. ochraceus* populations in Oregon appear relatively consistant over the time frame, while Alaska populations become more variable in 2014 as compared to the previous year, then decrease significantly by 2016 (fig 13a)

Next, time points were subdivided from year to sampling season in an attempt to resolve trends at a finer scale, then plotted as before (with E. troschelii excluded, fig 14). It is worth noting that *K. tunicata* is not an echinoderm, and therefore was wholly unaffected by SSWD, so it serves as a usefull baseline to compare *P. ochraceus* abundance trends during the SSWD event to. Indeed, *P. ochraceus* numbers drop from fall of 2013 to the spring of 2014, while *K. tunicata* levels remain consistant. These graphs display evidence that SSWD substantially altered *P. ochraceus* abundance following the start of the event in late 2013, as both the median of the distribution drops, and the majority of points (box of the boxplot) approach zero (fig 14)

### *Question 3: How did populations of the focal sp., P. ochraceus, respond & recover from the SSWD event?*

From here on, I focused exclusively on my focal species of interest, *P. ochraceus*, and moved towards a slightly different approach to visualize changes before & after the focal SSWD event by using density plots with quartile divisions. Further, I narrowed my scope of investigation by filtering only sites which had high abundance counts prior to the SSWD event (50+ individuals counted during 2013), as healthy, robust *P. ochraceus* populations are arguably of greatest ecological interest, as well as the general principle in disease ecology that infectious diseases are likely to spread more quickly in dense populations.

Figure 15 displays the abundance distributions over time for the six sites which met the above criteria, and the results are apparent: from 2013 to 2016, each site’s abundance distribution clearly shifted towards the left, meaning more observations had low (or even zero) counts. Sites “5030” and “6220” display similar and extremely intriguing trends: both sites have relatively normal abundance distributions in 2013 that center around 3 (log(Abundance)) (fig 15). In 2014, this distribution is disrupted and shifts left. However, in the last two years displayed, both sites appear to develop a bimodal distribution, suggesting potential recovery (fig 15).

Evidence of population recovery may imply that young recruits are beginning to recolonize an area previously disrupted, so I was interested in seeing if there were any detectable shifts in size distributions of *P. ochraceus* individuals that corresponded to explainable trends in abundance. In order to give historical context to size distributions, the range of years was expanded out again from 2010 to 2018. Figure 17 gives a side-by-side comparison of abundance and size class distributions for *P. ochraceus* during this time frame across all sites surveyed, and figure 18 displays the same information, but only for sites characterized by a high abundance in 2013 (50+ individuals). Across all sites, the growing abundance distribution peak near/at zero reflects trends seen in earlier plots, and it is interesting to note the more subtle, but present, shift of the distribution (most easily viewed by focusing on relative quartile position) to the left from 2013 to 2014, then the gradual shift back towards the right from 2015 onward (fig 16). This is even more apparent when viewing the mobility of the third and fourth quartiles (two darkest shades of blue) in filtered site plot (filtered to only include high count sites; fig 17).

Finally, I wanted to visualize this information in another way that highlighted extreme values, to demonstrate the idea that some sites are recovering, as evident by observations that have high abundance measurements and small size classes. Figure 18 does exactly that, by plotting abundance (raw, fig 18a, and log-transformed, fig 18b) over time with points colored based on their size class. It is strikingly apparent that while there are, indeed, more observations with high abundance values after around 2014, these observations feature much smaller sea stars (as shown by the purple-shading of these points, corresponding to a lower size class, fig 18a-b). Raw abundances were included (fig 18a) was included to emphasize outlier sites, and the regression line (model = LOESS, locally weighted smoothing) of the log-transformed abundances (fig 18b) helps to visually describe the average statistical trend.

## Full analysis

### *Machine Learning Overview*

The machine learning component of this study consisted of two main analyses, performed on a subset of the data which only included entries on the focal species, *Pisaster ochraceus*. The first used variable selection to address the question *“Which variables are good predictors of species abundance?”*; the second explored the data using tree modeling to ask *“Are there definable geographical regions based on the available predictor variables?”*. Both analyses employed a data splitting strategy, where 70% of the data (“train set”) was used to train and select the best-fitting model, and the remaining 30% (“test set”) was retained for use in the last step of the procedure, which evaluated the performance of the final model on the test set. The script for the variable selection analysis is found here: ./code/analysis\_code/Continuous\_Outcome\_Modeling.Rmd, and the script for the tree fitting analysis is located here: ./code/analysis\_code/Categorical\_Outcome\_Modeling.Rmd.

### *Question 1: Which variables are good predictors of species abundance?* ***Variable Selection***

I used statistical models to assess the relationship between predictor variables and an outcome of interest, *P. ochraceus* abundance (the variable total) , using a variable selection approach.

Figures 19 and 20 display the bivariate relationships between the numeric and categorial predictors (respectively) and the outcome of interest, ochre sea star (*P. ochraceus*) abundance (log-transformed “total”). Some variables were then were removed due to their redundancy, and a subset of 19 predictors were used in the variable selection analysis: “site\_code”, “georegion”, “mpa\_region”, “bioregion”, “size\_bin”, “group\_code”, “group\_code\_UCSC\_other”, “marine\_season\_code”, “longitude”, “season\_name”, “latitude”, “method\_code\_IP\_other”, “state”, “marine\_sort\_order”, “method\_code”, “island”, “marine\_common\_season”, “marine\_common\_year”, and “species\_code”.

First, single predictor models were run on all 19 potential correlating factors and evaluated using the performance metric RMSE (root mean squared error). All 19 predictors outperformed the null model (null model RMSE = 17.13, red dashed line in fig 21). Further, the predictor “site\_code” stood out as having the lowest RMSE by a fair margin; “georegion”, “mpa\_region” and “bioregion” also performed well (fig 21). This suggests that predictors describing locations, and especially those describing smaller ranges (as opposed to broader spatial descriptors, like “longitude”, “latitude”, and “state”), are good predictors of *P. ochraceus* abundance.

Next, fitting with multiple predictors was performed, using three different models: linear (lm), regression splines (mars), and K nearest neighbor (knn). Further these models were run with and without a preprocessing step (centering and scaling, denoted by "\_cs" suffix, fig 23). Interestingly, the centering and scaling worsened the performance of the best-fitting model, knn, though all outperformed the null model (null model RMSE = 17.13, red dashed line in fig 21).

Finally, after the knn model residuals were plotted and no abberent patterns found (fig 24), I validated the results by assessing of the final model (knn) on the test set of data. As expected, both knn models (on the training and test data, darker shaded boxes, fig 25) outperformed the null models (lighter shaded boxes, fig 25). Somewhat surprisingly, however, the model performed even better on the test data than it did on the training data. This gives evidence to support the notion that the model is not overfitted to the training data, and suggests the final model will be robust to new data.

### *Question 2: Are there definable geographical regions based on the available predictor variables?* ***Tree Fitting***

The second machine learning component focused on the categorical variable bioregion and used variable selection and a tree fitting method to attempt to assign geographical region to an observation given predictor variables. The categorical outcome of interest, bioregion, consisted of six biogeographical subsections (from north to south): Alaska, USA to British Columbia, Canada; Salish Sea, WA, USA; Olympic Coast, WA, USA to San Francisco, CA, USA; San Francisco to Government Point, CA, USA; Channel Islands South, CA, USA (west of Los Angeles); Government Point, CA, USA to Mexico. The majority of observations were in two adjacent regions: Olympic Coast to San Francisco and San Francisco to Government Point (fig 26).

Figures 27 and 28 display the bivariate relationships between the categorial and numeric predictors (respectively) and the outcome of interest, bioregion; these variables were then used in the single predictor models, which evaluated the training data using the performance metric accuracy (ACC), which represents the overall agreement rate averaged over cross-validation iterations and is displayed as a percentage/proportion (fig 29). The null model for accuracy, which simply assigns all observations in the most populous category was 42.6% (red line, fig 29), was outperformed by all single predictor models (fig 29).

At this stage, it was noted that several predictors had very strong (or even perfect) correlations with biogeography, with very logical explanations as to why this was the case (the six most highly correlated were directly related to location, fig 29). Therefore, predictors with over 70% accuracy at this stage were removed for all subsequent analyses (following this filtering step, the data was resplit into test/train groups).

Next, fitting with multiple predictors was performed, using five different models: logistic (tree), GBM (boosted tree ensemble), PDA (discriminant analysis), RF1 (random forest), and RF2 (random forest with pre-processed predictors). Interestingly, the tree model performed far worse than the other four multiple predictor models (ACC = 0.4), so it was not plotted in figure 30 (and the decision tree was not included as a figure). Based on the performance metric accuracy and kappa (average Cohen’s unweighted Kappa statistic of resampling results), the boosted tree ensemble (GBM) model predicted bioregion the best, so it was selected for the final analysis. This model uses a collection (or “ensemble”) of weak learner trees that perform poorly individually but often very successfully in combination, so it is not surprising that this model predicts the outcome well.

Figures 30 and 31 display confusion matrices, which describe the performance of the boosted tree ensemble (GBM) model on the subset training data and the subset test data, respectively. Further, both GBM models (on the training and test data, darker shaded boxes, fig 33) outperformed the null models (lighter shaded boxes, fig 33). This model performed almost exactly as well on the test data as it did on the training data, suggesting that, in addition to having 96% accuracy, it is not overfitted to the training data and would be robust to new data. This is especially desirable when the goal of the model-fitting is to predict with high accuracy.

It is not surprising that when bioregion is incorrectly assigned, it is to another bioregion that is proximal to it geographically. Further, the two categories most often mis-assigned to as the other are “Olympic Coast to San Francisco” and “San Francisco to Government Point” (figs 30-31). While it is possible this is an artifact of their being the two most sampled bioregions, it is possible that this may be indicative of the idea that these two regions are most similar to one another, at least for a sea star.

## Discussion

### *Summary and Interpretation*

#### Abundance Trends

A primary aim of this investigation was to make use of the extensive monitoring data available through MARINe to investigate spatial and temoporal changes in target species assemblages along the west coast of the United States. Although the mottled sea star (*Evasterias troschelii*) observations were too limited to fully assess abundance changes over time, the data yielded intriguing trends regarding the other two species, the black chiton (*Katherina tunicata*) and the ochre sea star (*Pisaster ochraceus*). While *K. tunicata* abundance measures were fairly uniform across time and space, *P. ochraceus* varied considerably over time, for reasons which were answered in the second part of the exploratory analysis.

The overarching goal of this investigation was to better understand how *P. ochraceus* populations changed in the years surrounding a catastrophic sea star wasting disease (SSWD) outbreak, which occured from 2013-2014. Indeed, the univariate and multivariate approaches I used to visualize trends in ochre star abundance suggested that populations in the Pacific northwest varied over time in a manner that was consistent with SSWD reports. I observed discernable changes in spp. abundance before & after the focal SSWD event; more specifically, there was a marked period of reduced abundance around 2014, which is evident across all sites but especially in those sites which had large starting populations before the disease event. This observation supports the notion that infectious diseases generally spread more quickly in dense populations.

Fortunately, there was also evidence to suggest that some *P. ochraceus* populations are recovering in the years after the outbreak occured. Perhaps the most interesting result of this analysis was the observation that, not only was there a clear we see a decrease in abundance at the expected time point (2013-2014), but there were also a handful of sites which had spikes in abundance later on (2015-2016). The observation that sites with high star counts during this time also featured unusually small stars generates a compelling trend, implying that some populations are recruiting offspring and, thereby, recovering from near-catastrophic mortality rates.

#### Predictive Relationships

The two machine-learning components of this study focused exclusively on *P. ochraceus* observations; both generated models that vastly outperformed expectations based on the null models and were highly robust to “new” data (ie, test data from a 70:30 train:test split approach). The single-predictor modeling portion of the variable selection analysis (with *P. ochraceus* abundance as the outcome of interest) revealed that small-scale spatial predictors like bioregion, MPA region, georegion, and site location performed well in predicting species abundance. This suggests either that population size is variable across the *P. ochraceus* range or that population size is very consistant over time within a given site or region; however, the latter seems highly unlikely given the temporal trends observed in the exploratory analyses.

The second machine-learning exercise trained models to predict bioregion based on the data provided and the very high success rate of the final model (>95% correct assignment) on both the training and test data suggests either one of two things: (1) the six bioregion classifcations used to characterize the west coast in this survey are highly divergent from one another and accurately deliminate important factors which, in turn, heavily influence *P. ochraceus* populations or (2) even after removing a substantial portion of highly correlated variables, the predictors used in the model were “too descriptive” and made it easy to predict bioregion. Unfortuntely, I suspect it is the latter.

### *Strengths and Limitations*

As alluded to in the previous section, many of the predictors were highly correlated with the outcome of interest in the second machine-learning exercise, so the high success of the second modeling activity is largely uninteresting because highly-correlated variables should be good predictors of the outcome they are correlated with.

Another percieved weakness of the machine-learning portion of this study was that, although I employed the test/train split strategy, the gold-standard is to perform cross-validation, whereby

Finally, the exploratory analysis portion of this investigation relied on visually perceptible trends. The conclusions I’ve drawn would be more robust with statistical testing, but I opted to dedicate my efforts elsewhere to produce a compelling investigation. If I were to publish this work, I would statistically validate the differences I observed and described herein.

I believe a strength of this study was the effort made to vary both temporal and spatial scales at which trends were analyzed, as well as the well-thought-out and highly deliberate usage of different visualization tools to highlight different trends. This was very time-consuming, but incredibly useful in not only revealing trends that would be imperceptible when viewed at a lower resolution, but also in refining my skills and increasing my confidence as an R user.

### *Conclusions*

Ochre sea star populations suffered high rates of loss during the 2013-2014 sea star wasting disease event on the west coast of the US, but, herein, I provide evidence that they are capable of recovering, at least under some situations. This suggests future efforts should be concentrated on elucidating what distinguishes these “high-abundance small-size-class” sites from those whose numbers do not appear to be recovering.

### *References*

Bucci, C., Francoeur, M., McGreal, J., Smolowitz, R., Zazueta-Novoa, V., Wessel, G. M., & Gomez-Chiarri, M. (2017). Sea Star Wasting Disease in Asterias forbesi along the Atlantic Coast of North America. *PLOS ONE*, *12*(12), 1–20. <https://doi.org/10.1371/journal.pone.0188523>

Burge, C. A., Mark Eakin, C., Friedman, C. S., Froelich, B., Hershberger, P. K., Hofmann, E. E., … Harvell, C. D. (2014). Climate Change Influences on Marine Infectious Diseases: Implications for Management and Society. *Annual Review of Marine Science*, *6*(1), 249–277. <https://doi.org/10.1146/annurev-marine-010213-135029>

Eisenlord, M., Groner, M., Yoshioka, R. M., Elliott, J., Maynard, J., Fradkin, S., … Harvell, C. D. (2016). Ochre star mortality during the 2014 wasting disease epizootic: Role of population size structure and temperature . *Philosophical Transactions of the Royal Society B*, 201502212. <https://doi.org/https://doi.org/10.1098/rstb.2015.0212>

Gudenkauf, B. M., & Hewson, I. (2015). Metatranscriptomic Analysis of Pycnopodia helianthoides (Asteroidea) Affected by Sea Star Wasting Disease. *PLOS ONE*, *10*(5), 1–10. <https://doi.org/10.1371/journal.pone.0128150>

Harley, C. D. G., Pankey, M. S., Wares, J. P., Grosberg, R. K., & Wonham, M. J. (2006). Color Polymorphism and Genetic Structure in the Sea Star Pisaster ochraceus. *The Biological Bulletin*, *211*(3), 248–262. <https://doi.org/10.2307/4134547>

Hewson, I., Button, J. B., Gudenkauf, B. M., Miner, B., Newton, A. L., Gaydos, J. K., … Harvell, C. D. (2014). Densovirus associated with sea-star wasting disease and mass mortality. *Proceedings of the National Academy of Sciences*, *111*(48), 17278–17283. <https://doi.org/10.1073/pnas.1416625111>

Kohl, W. T., McClure, T. I., & Miner, B. G. (2016). Decreased Temperature Facilitates Short-Term Sea Star Wasting Disease Survival in the Keystone Intertidal Sea Star Pisaster ochraceus. *PLOS ONE*, *11*(4), 1–9. <https://doi.org/10.1371/journal.pone.0153670>

Lessios, H. A. (1988). Mass Mortality of Diadema Antillarum in the Caribbean: What Have We Learned? *Annual Review of Ecology and Systematics*, *19*, 371–393. Retrieved from <http://www.jstor.org/stable/2097159>

Miner, C. M., Burnaford, J. L., Ambrose, R. F., Antrim, L., Bohlmann, H., Blanchette, C. A., … Raimondi, P. T. (2018). Large-scale impacts of sea star wasting disease (SSWD) on intertidal sea stars and implications for recovery. *PLOS ONE*, *13*(3), 1–21. <https://doi.org/10.1371/journal.pone.0192870>

Montecino-Latorre, D., Eisenlord, M. E., Turner, M., Yoshioka, R., Harvell, C. D., Pattengill-Semmens, C. V., … Gaydos, J. K. (2016). Devastating Transboundary Impacts of Sea Star Wasting Disease on Subtidal Asteroids. *PLOS ONE*, *11*(10), 1–21. <https://doi.org/10.1371/journal.pone.0163190>

Orth, R., Carruthers, T., Dennison, W., Duarte, C., Fourqurean, J., JR, K., … Williams, S. (2006). A Global Crisis for Seagrass Ecosystems. *BioScience*, *56*, 987–996. <https://doi.org/10.1641/0006-3568(2006)56[987:AGCFSE]2.0.CO;2>

Schiebelhut, L. M., Puritz, J. B., & Dawson, M. N. (2018). Decimation by sea star wasting disease and rapid genetic change in a keystone species, Pisaster ochraceus. *Proceedings of the National Academy of Sciences*, *115*(27), 7069–7074. <https://doi.org/10.1073/pnas.1800285115>

Schultz, J. A., Cloutier, R., & Côté, I. M. (2016). Evidence for a trophic cascade on rocky reefs following sea star mass mortality in British Columbia. *PeerJ*.

Waycott, M., Duarte, C. M., Carruthers, T. J. B., Orth, R. J., Dennison, W. C., Olyarnik, S., … Williams, S. L. (2009). Accelerating loss of seagrasses across the globe threatens coastal ecosystems. *Proceedings of the National Academy of Sciences*, *106*(30), 12377–12381. <https://doi.org/10.1073/pnas.0905620106>