SOAR monitoring Data Analysis

Jordanna Barley

1/22/2023

# Intoduction

The purpose of this document is to go over the exploratory data analysis that I have done for the SOAR monitoring data. First, exploratory data analysis is done to visualize the data and ensure that there does not seem to be any mistakes that would result from incorrect data entry. Second, this step allows you to start to recognize patterns in the variation in the data, which can sometimes allude to significant differences amoung treatments when statistical models are run. Third, this step allows you to make decisions about the type of model that you should run based on the observed patterns in the data visualization. Although you should already have a good idea of what types of models you are going to run based on the type of data that you collected, it is important to look at what the error/variation in your data look like, as this can change how you account for this error in your statistical model. Therefore, I have included plots and graphs to visualize the data collected from all monitoring sites that collected usable data in a variety of helpful combinations so that we can think about what to do going forward with this analysis.

## Survival

First, I looked at survival of the SOAR oysters deployed. I calculated proportional survival as the proportion of live oysters collected to the total number of oysters collected during the monitoring. Except for the two sites in New York which deployed bags of oysters, we can’t really calculate actual survival of SOAR oysters from each site. This is because we can’t say for sure what percentage of SOAR oysters were found/collected during monitoring in many sites because they deployed oysters at already estabolished oyster restoration sites. Therefore, I have calucalted the proportion of live oysters to total oysters as a proxy to show the relative abundance of live to dead oysters at each site. This way, proportional survival can be compared across sites and states no matter the deployment and sampling protocol.

Here, we are looking for any differences between sites, collection agencies, and states. It is really important to know if there are survival differences so we can be informed when thinking about sites for SOAR 2.0. Also, and maybe more importantly, survival is one of the metrics to look at for the efficacy of the SOAR program. If survival was generally high across the board, we can confidently say that this approach should be successful in most restoration sites. However, if there is some variation in survival between sites or states, then we might need to be more strategic about what sites are chosen and be more dilligent about the data that are collected.

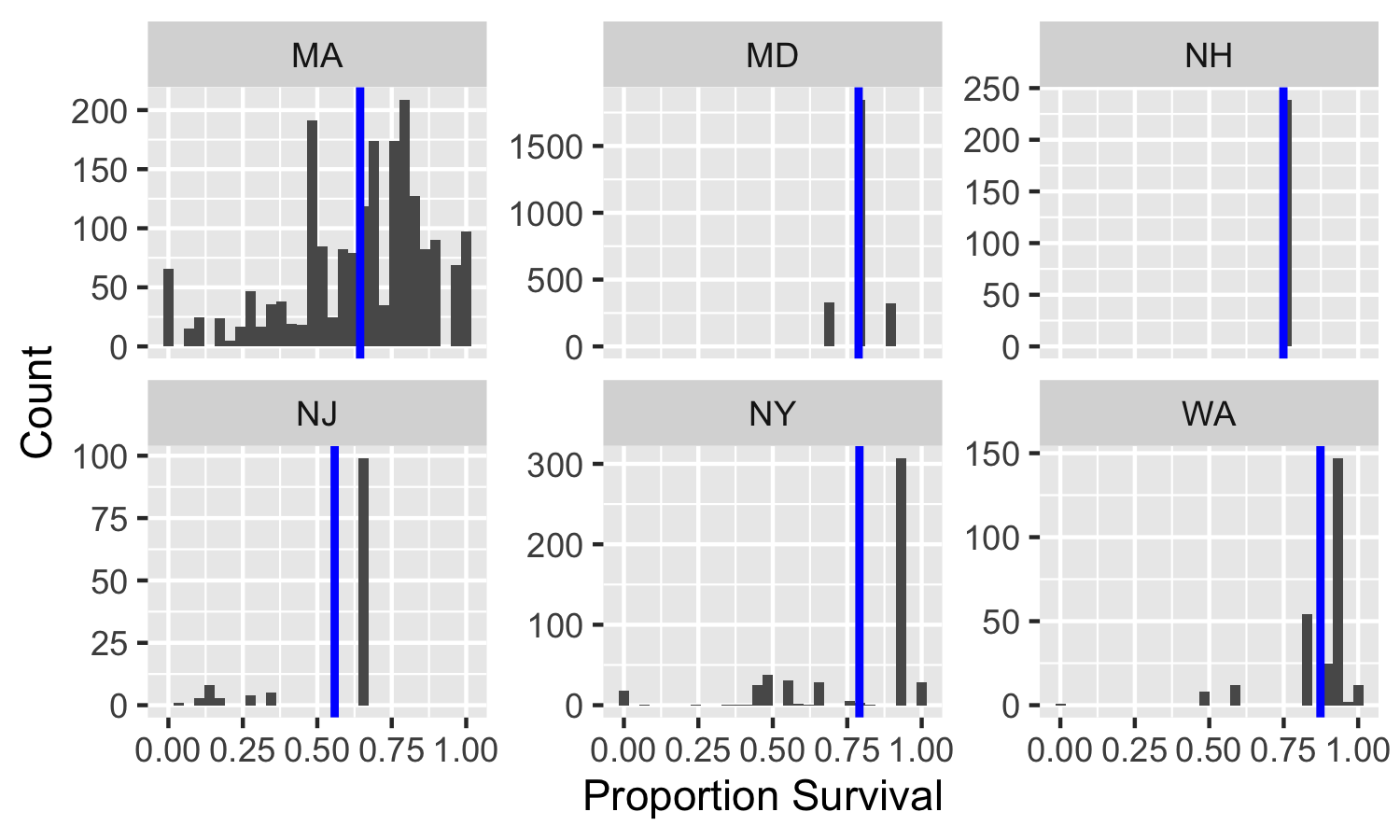


Fig. 1. Histograms of Proportional survival by state. Blue line denotes mean.

In Figure 1, you can see the distribution of proportional survival for each state. Most states don’t seem to have very much variation, except for Massachusetts. While it seems that most sits and replicates had gerater than 50% survival, this is in contrast to the other states that had relatively much smaller distributions and a mean of proportional survival closer to 75%.

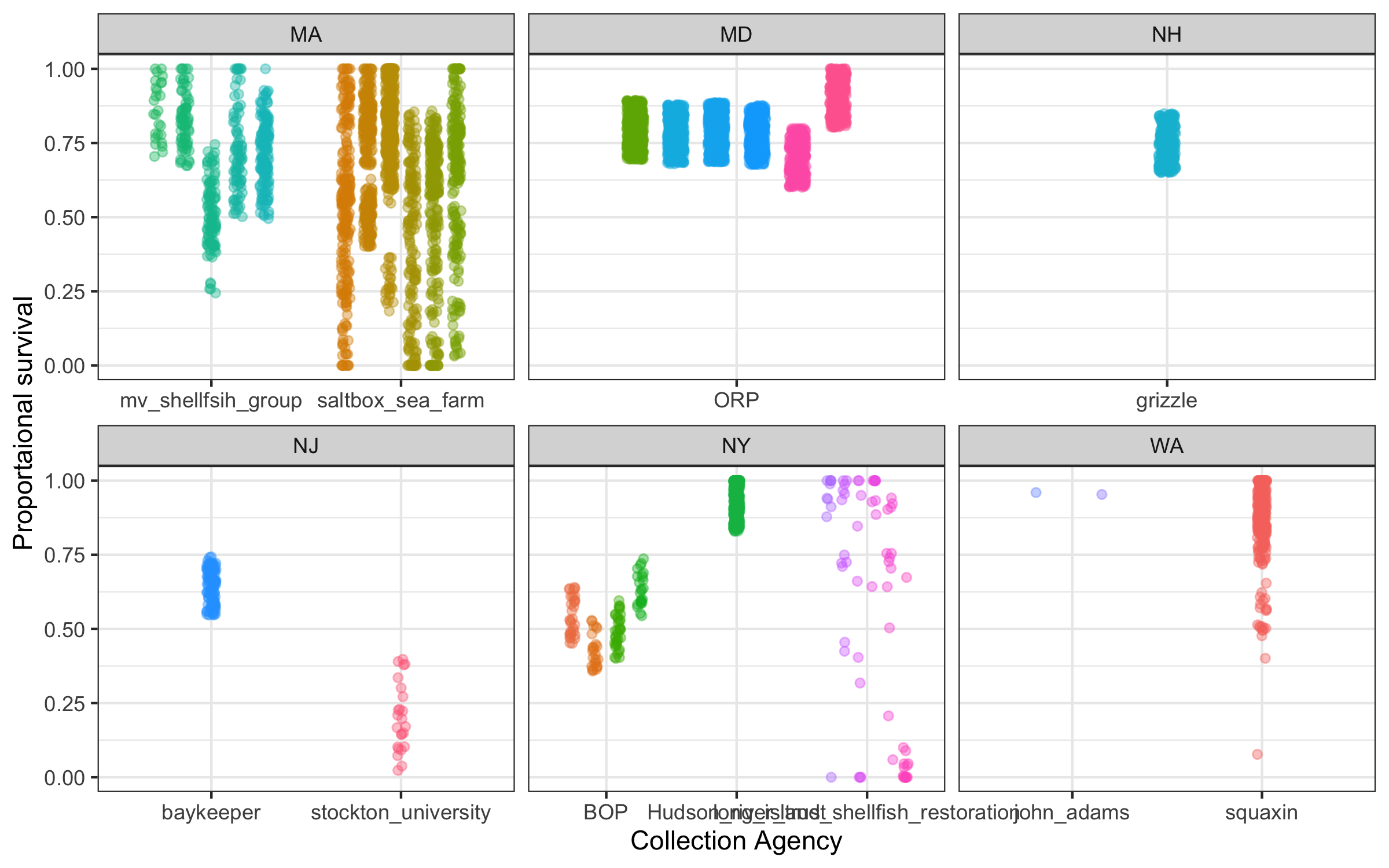


Fig. 2. String plots of proportional survival by collection agency and state. Different color data points denote sites within each collection agency.

If we break proportional survival down further by collection agency and site (Fig. 2), we can see more specifically that even some sites within each collection agency had more variation than others. This is to be expected - not only are the data collected by different people, in different states and sites - each site can vary largely in temperature, salinity, water flow velocity, predation, disease prevelance, etc.

### Preliminary Survival Statistics

I ran some prelininary models with the survival data. Because of the type of data and the dependent variables we have (state, collection agency, and restoration site), I ran several models with different combinations of the dependent variables being included. This is done because we want to make sure that the model that we use sufficiently explains the variation in our data well. I then use a technique to analyze which model explains the most variation in the data. This essentially means that I am looking for the best-fit model. You can think of putting a line of ‘best fit’ on a graph in excel as an example of what I am doing with the monitoring data. I have provided more details of the methods that I used at the end of this report.

The dependent variables that we are using are all highly collinear with each other. This means that the variation that we see in survival by state, collection agency, and restoration site are too similar. When you run a model that includes very collinear variables together, it is hard to trust if it is the ‘best fit’ because this could be from the collinear dependent variables OR that the model explains the variation in the data sufficiently. Therefore, we want to minimize the collinearity of variables included in the model. The models that best fit our data all include at least two out of the three dependent variables, making these models hard to interpret. I will say, however, that including the restoration site in the model seems to be important. Although this does not mean there are significant differences between each and every site, it might mean that survival varys across sites enough to show that this is an important factor in the survival of SOAR oysters. Though because of the notes on collinearity, I would take this last thought with a grain of salt.

## Growth

Next, I calculated growth for the states that provided pre-deployment size data. Growth was calculated by averaging the pre-deployment size data and subtracting that from the sizes of the individual oysters from each site collected for monitoring. Because we are not able to track the size of each individual oyster over time, these data are a bit messy. For example, there are many growth data points that are negative because that individual was measured at smaller than the average pre-deployment size for that site. We can think through how to interpret these data together when we meet; I think it can be informative to look at the distribution of growth across sites and states as well as what the mean is. For example, if the mean is around zero, we might interpret that as very little growth overall, even though it is unlikely there was no growth at all. We can also look at the mean size of pre-deployment and post-deployent oysters and compare the distributions of the two. This might end up being more informative but I figured it would be best to provide both so that we can discuss both options.

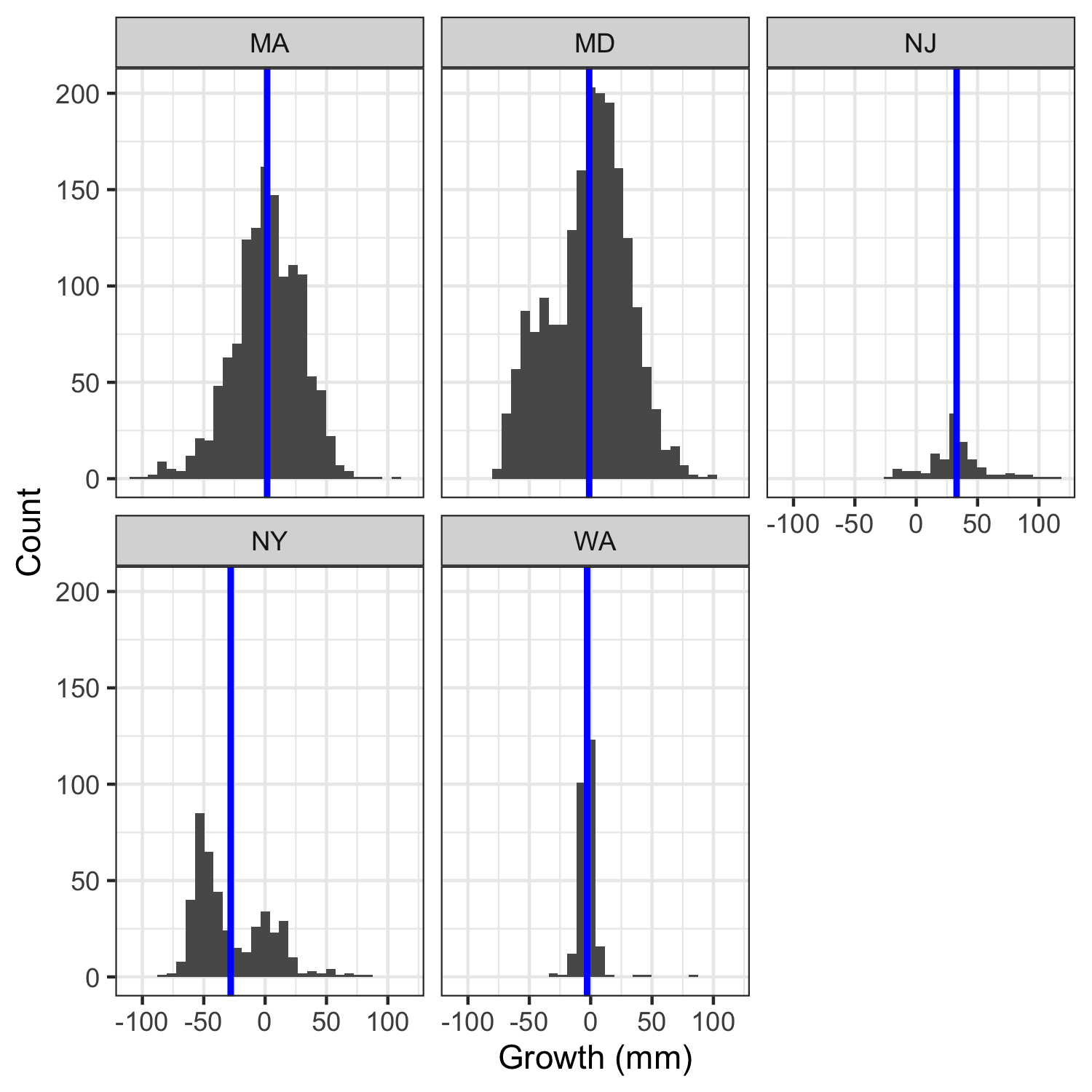


Fig. 3. Histograms of growth by state. Growth was calculcated as the the post-deployment size minus the average pre-deployment size for each site. Blue line denotes mean growth.

Washington, Maryland and Massachusetts oyster growth is centered around zero- as I mentioned before, we could interpret this metric as no average growth for that site (Fig. 3). New Jersey seems to be skewed in the positive direction, which could mean overall growth for the sites in the state. Lastly, Ney York growth has a bimodal distribution, it is possible that one site is dragging the mean for the entire state down.

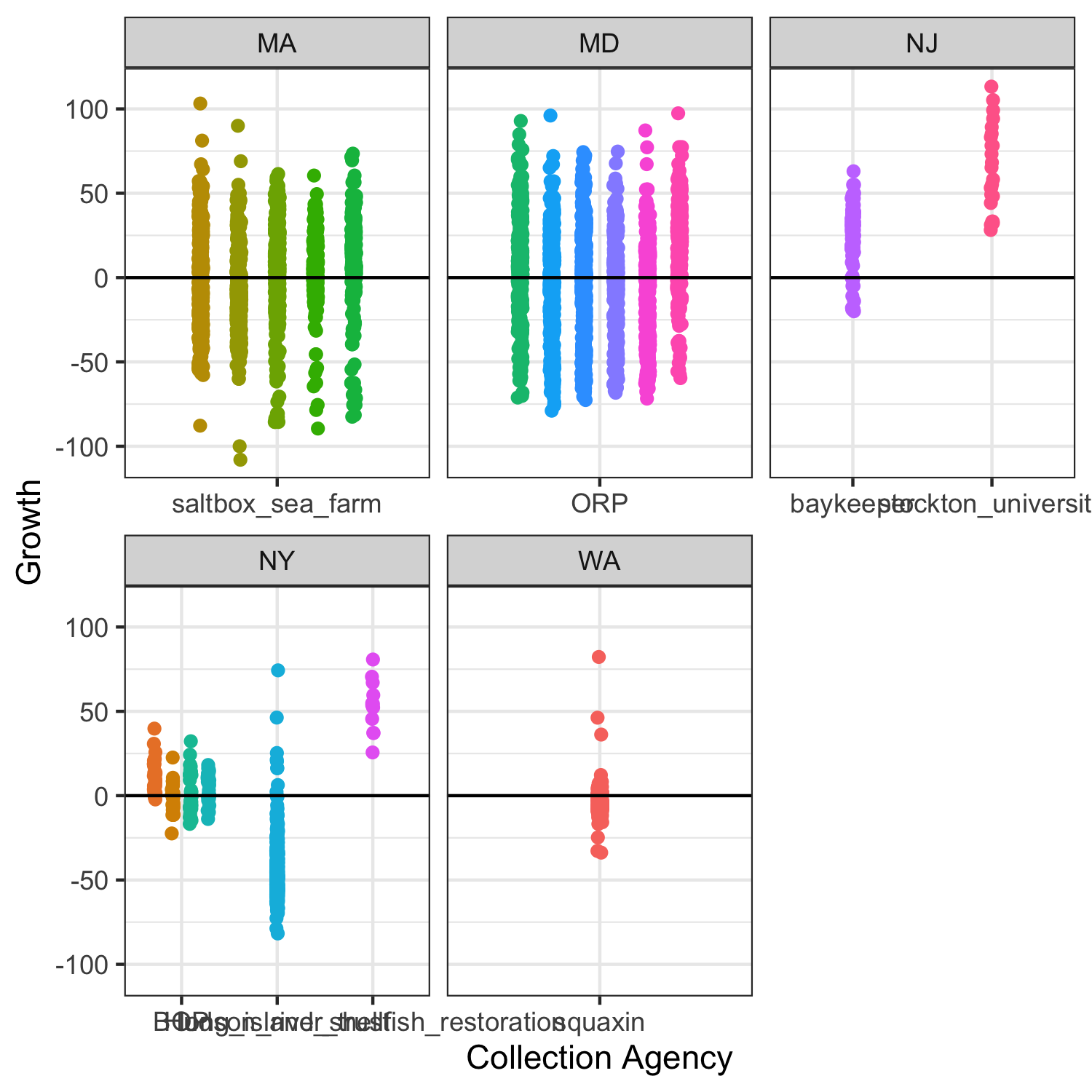


Fig. 4. String plots of growth by state and collection agency. Different colors denote different restoration sites and the black lines are at zero.

It is clear that the Hudson River Trust sites are what is causing the bimodal distribution that is seen in Fig. 3 (Fig. 4).

I meant to evaluate growth by testing the differences in the average sizes of pre-deplyment and post-deployment but I ran out of time. Because of that, I have not completed the statistical analysis of growth yet but should be able to by the next time we mee!

## More detailed statistical methods

For survival, I ran generalized linear models with a binomial error distribution. I ran one model for each single dependent variable on its own (state, collection agency, restoration site) in addition to additive models for each combination of the variables. I ran a Variation Inflation Factor (VIF) that basically said this dependent variables are very collinear. I also compared all of the models that I ran using Akiake Information Criterion (AIC). Three of the models had the exact same corrected AIC value and collectively they were weighted at 90% for model selection. A fourth had a very similar corrected AIC, and was about 2 delta AIC units away from the other three models. All four of these models included restoration site as an independent variable, which is way I included the note above about variation between sites.

Glm(proportional survival~ state+ collection agency+ restoration site, family=binomial)