Investigating Regional Effects on North American Pacific Salmon Survival

Peter Jacobs

4/30/2018

1: Introduction

1.1: Question Description and Motivation for Analysis

The health and longevity of the North American Pacific salmon population is vitally important for our appetites and economy. The 2015 National Oceanic and Atmospheric Administration (NOAA) Fisheries of the United States report states that U.S fisherman landed a salmon haul worth a value of 460.2 million dollars in 2015. Also according to the report, a total of 70% of all fish landings in 2015 came from the Pacific and Alaskan coasts. This brings to mind two important questions regarding Salmon populations along the North American pacific coastline

- Is the region in which salmon stock spawn associated with differences in the health of these salmon stocks?
- If there is such an association, what are the underlying region level variables that dictate [not necessarily in a casual way] the magnitude of this association (e.g. the weather in the region? The fresh water quality in rivers in the region? The amount of human development in the region?)

The subsequent analysis in this report attempts to begin to investigate these questions using data starting from 1950 on the spawning and recruiting counts for North American salmon stocks from different regions of the pacific coast.

1.2: Exploratory Analysis

The purpose of the exploratory analysis is to try to begin to understand how region is associated with the health of stock in a given year.

The first to thing to think about is how to measure the health of the stock. In this analysis, this is done by taking the number of spawners in a given brood year and dividing by the number of recruiters in the previous brood year. This will be called the "life cycle survival rate" (lcsr) of the stock in a given brood year.

Using the lcsr of the stock as a "response variable", the data is visualized to try to gain insight into the questions outlined earlier. Figure 1 [1] shows a map with the location of each stock population, where the size of point indicates the magnitude of average lcsr. From north to south, there doesn't appear to be any clear trend in the lcsr. Figure 2 shows the marginal histogram of the lcsr's. They seem to spread across [0,1] rather evenly (except there are very few extremely small values). Figure 3 breaks the lcsr down by region. Note that certain regions tend to have higher lcsr than others (e.g Yakutat has very high lcsr's, whereas most other regions have lcsr's across the board). The wide distribution of the lcsr probably has something to do with the variation in stock. Figure 4 shows figure 2, but broken down further by stock. Within region, there does appear to be some breakdown by stock. This seems clearest in the Norton Sound region. Also relevant is how the lcsr changes as the along shore distance changes. Figure 5 shows the average lcsr at each along-shore distance. It is hard to capture any noticeable pattern here.

Important takeaways from the entire exploratory data analysis are as follows:

[1] All figures and tables in this report are included in the appendix

- There does appear to be some difference in the lcsr based on region.
- Within each region, there are a variety of factors that are probably influencing lcsr (e.g time and stock population)
- It does't seem like along shore distance is associated with lcsr

2: Methods

The strategy for model building was to build successively more complex models, and in the process, check for convergence using trace plots. Each model is fit using all rows in the main dataset that have no missing values for the variables that are used in the model.

The algorithm settings and trace plots for models 1,2 and 3 are reserved for the appendix. Also, the interpretations and assumptions of models 1 and 2 (the non final models) are in included in the appendix, while for model 3, this information is in this section.

2.1: Model 1

The first model is the simplest of all models.

Likelihood:

$s_{ij}|r_{ij}, p_{ij} \sim binom(r_{ij}, p_{ij})$ for $j = 1, ..., R$ and $i = 1, ..., n_j$ where R is the number of regions in the dataset, and n_j represents the number of broodings involving stocks from region j. s_{ij} represents the number of spawners from the i^{th} brooding in region j. r_{ij} represents the number of recruiters from the previous brooding of the stock that gives rise to the i^{th} brooding in region j; p_{ij} represents the probability that a salmon that was a recruiter in the stock that gives rise to the i^{th} brood in region j will be a spawner in the i^{th} brooding in region j.

$logit(p_{ij}|\alpha_j) = \alpha_j$

Prior:

$\alpha_j \sim N(0, 1000)$

Motivation for Next Model

As can be seen in figure 6, it appears that the posterior samples for this model converge. Moreover, the α_j in this model could be used to attempt to understand how region is associated with the survival of salmon stocks. However, there are so many other variables that contribute to the ability of salmon populations to flourish, that it is probably better to include some of these factors in the model. For example, there is probably systematic variation in the ability of salmon stocks to survive from recruitment to spawning over time. Therefore, it might be useful to include time in the model. If this is done, then the interpretation of regional effects can be done conditional on a given time. Another consideration is that there is probably stock-brood by stock-brood variation in the logit probability of survival, so it may be a good idea to include an overdispersion parameter in the model. Finally, while model 1 may help us understand the marginal effect of region, we are also interested in the underlying variables that drive the generating process for the regional effects. For example, does the median along shore distance of brooding areas within a given region contribute to the magnitude of the regional effect? The next model will try to address this question, while also including an overdispersion parameter, and accounting for the year in which a stock breeds.

2.2: Model 2

Likelihood:

The likelihood begins with the first paragraph from model 1 (see model 1 likelihood)

$logit(p_{ij}|\alpha_j, \beta_0, \beta_1, \gamma_i) = \beta_0 + \beta_1 * t_{ij} + \alpha_j + \gamma_i$ where t_{ij} represents the year in which the i^{th} stock from region j performed breeding.

$\gamma_i|\tau^2 \sim N(0, \tau^2)$

$\alpha_j|\beta_2, \tau_\alpha^2 \sim N(asd_j * \beta_2, \tau_\alpha^2)$ where asd_j represents the median along shore distance of stock breeding sites within region j.

Prior:

The following vague priors are used:

$\beta_0 \sim N(0, 1000)$ $\beta_1 \sim N(0, 1000)$ $\beta_2 \sim N(0, 1000)$ $\tau^2 \sim invGamma(.001, .001)$ $\tau_\alpha^2 \sim invGamma(.001, .001)$

Motivation for Final Model

Model 2 incorporates the potential dependence of survival probability for a stock on both spatial and temporal variables (which seems more realistic a scenario than model 1). By analyzing the α_j, we can analyze whether there are regional effects on stock survival rates. By analyzing β_2, we can begin to understand where the α_j come from (but not nescessarily in a casual sense). This will help us answer the second question posed in the introduction of this report. However, as can be seen in figure 7, using JAGS, the MCMC sampler does not converge (even with 50,000 adapt steps, 25,000 burn in iterations, and 25,000 saved steps). For this reason, this model will not be the final model used for analysis in the results section. Model 3 is a stripped down version of model 2 that both allows us to try to answer the questions set forth in the introduction (at least partially), while also having MCMC samples that converge.

2.3: Model 3 (Final Model)

Likelihood:

The likelihood begins with the first paragraph from model 1 (see model 1 likelihood)

$logit(p_{ij}|\alpha_j) = \alpha_j$
$\alpha_j|\beta_2, \tau_\alpha^2 \sim N(asd_j * \beta_2, \tau_\alpha^2)$ where asd_j represents the median along shore distance of stock breeding sites within region j.

Prior: The following vague priors are used:

$\beta_2 \sim N(0, 1000)$ $\tau_\alpha^2 \sim invGamma(.001, .001)$

Assumptions:

- Conditional independence of the spawn counts for a given stock brooding given the recruit counts for that stock brooding (from the previous brooding of that stock) and life cycle survival probability for that stock brooding
- Conditional independence of the regional effects given β_2 and τ_α^2

Interpretations:

α_j: The effect of brooding in region j on the logit probability of survival (from recruitment to spawning) in the breeding of a stock in region j.

β_2: For a one unit increase in the median along shore distance of breeding sites within a region, there is a β_2 unit increase in the effect of region on the logit life cycle survival probability for stock broods.

Shown in figure 8 in the appendix, the MCMC samples for this model appear to converge. Also, as discussed before, since this model allows the questions from the introduction to be explored, this model is the "final model", and is used for analysis. Note that even just adding an intercept to this model leads to the non-convergence issue, hence this model doesn't include an intercept. Also note that the priors in all of these

models are vague. This is because the designer of these models (me) is not an expert in North American Pacific Salmon fisheries; in order to be cautious, the modeler has decided to use vague priors.

3: Results

3.1 Addressing Question 1

Recall that the first question of interest was whether or not the region in which Salmon stock spawn is associated with the health of the stock. This question can be investigated by analyzing the α_j parameters. Because the α_j are the only effect making up the log odds of survival, I will interpret the α_j on the probability scale (rather than the log odds scale). The posterior samples shown below are for $\frac{e^{\alpha_j}}{1+e^{\alpha_j}}$. In this model, $\frac{e^{\alpha_j}}{1+e^{\alpha_j}} = p_{ij}$. Figure 9 gives the distribution of posterior samples for the probability of salmon stock survival within each region.

From this plot, we see that there is a clear effect of region on the probability of salmon stock survival. For example, in Norton Sound, the approximate posterior mean probability of salmon surviving and returning to breed is .8682 while in Prince William Sound, the approximate posterior mean probability of salmon surviving and returning to breed is .1854. Table 1 the 2.5th and 97.5th quantiles of the posterior samples for each region [2]. In Norton sound, the approximate 95% posterior credible interval for the probability of salmon surviving and returning to breed is [.8680,.8683]. For Prince William Sound, the 95% posterior credible interval is [.1853,.1855].

3.2 Addressing Question 2

The second question asked what the underlying region level variables are that are associated with the magnitude of the effect that region has on life cycle survival probabilities. One inherent characteristic of region is location. One might wonder whether how northerly or southerly a region is relates to the magnitude of the regional effect on life cycle stock survival probability. We can analyze β_2 to answer this question.

Recall the β_2 parameter gauges how the median along shore distance of a region influences the regional effect on life cycle survival probability. The posterior samples for β_2 are shown in figure 10. And a 95% posterior credible interval for β_2 is [-.00021,.00027]. The interval contains zero. There is not evidence to suggest that the median along shore distance of region is associated with the regional salmon stock life cycle survival probability.

4: Discussion

This section has three parts. The first part discusses positive qualities of this analysis, and analytical conclusions regarding the questions put forth in the introduction. The second part discusses limitations of this analysis. The third part discusses a different modeling idea to address the questions posed in the introduction.

4.1 Conclusions Regarding Questions

The models built in this report are simple and easy to understand. Moreover, there seems to be a definitive answer to first question posed in this report. There is an effect of region on the health of salmon stocks.

[2] The posterior variability in the α effects are mysteriously small; this could be for a variety of reasons, but it is possible that the prior variances were set too small

Finally, this report has begun to delve into where this effect is coming from. Based on the analysis of β_2, it doesn't look like the northerly/southerly aspect of a region is associated with the probability of salmon stock survival in that region.

4.2 Limitations

There are a variety of limitations of this analysis. A couple of them are mentioned below:

- The second question in this report inquires about the source of the regional effect on salmon stock survival rates. Few conclusions are drawn about what underlying variables are associated with this effect. For example, there may be important regional level variables such as average whether conditions, water temperature, or human population density that drive the regional effect.
- The models built in this report are overly simplistic, and the most realistic model fails to converge. It is important to adjust for the temporal effect that may exist and influence salmon stock survival. But the final model in this report does not include a time effect. Also, there is probably something inherently different about the survival behavior of each salmon stock population. This analysis does not include a stock level effect.

4.3 Another Modeling Strategy

In this analysis, the response variable is the probability that a salmon in a stock that is a recruiter in a given brood year, is able to comes back to its home river to spawn the following brood year. For a given stock, rather than use this probability as the response variable, it would be interesting to use the trajectory of this probability over time as the response function. This may allow for a more insightful analysis because one major goal of analyzing this data is to gain insight into the longevity and stability of salmon stock populations. Understanding what is nescessary to achieve longevity and stability in a population of any kind may require understanding what traits of a population are associated with changes in the trajectory of the survival of that population.

5: Appendix

5.1: Exploratory Data Analysis Visualizations

Figure 1: Exploring Stocks by their Mean Life Cycle Survival Rates

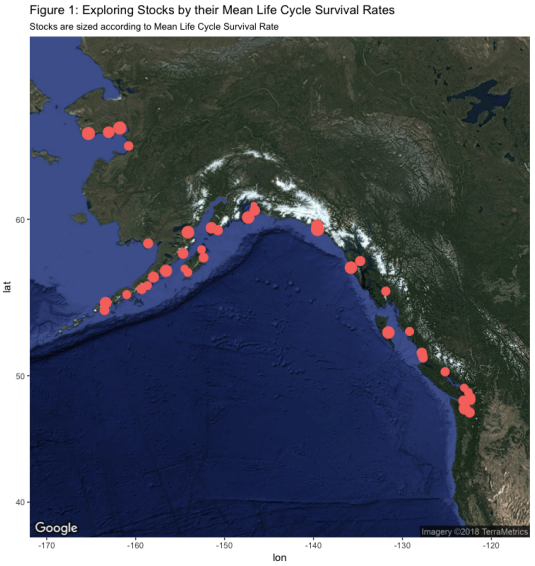


Figure 2: Life Cycle Survival Rate for Stock Breeding

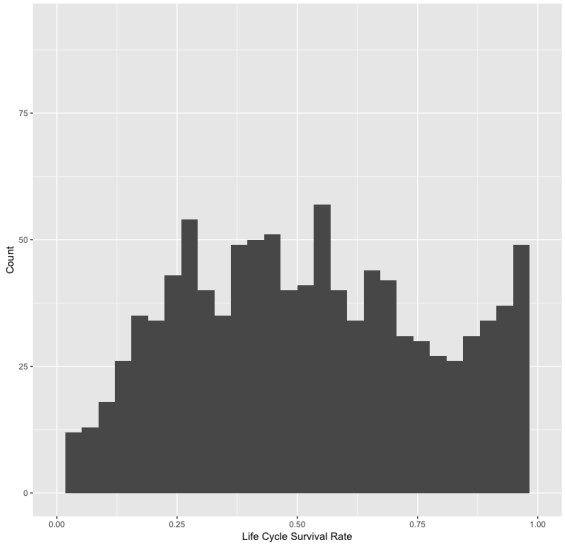


Figure 3: Life Cycle Survival Rate for Stock Breeding Broken Down By Region

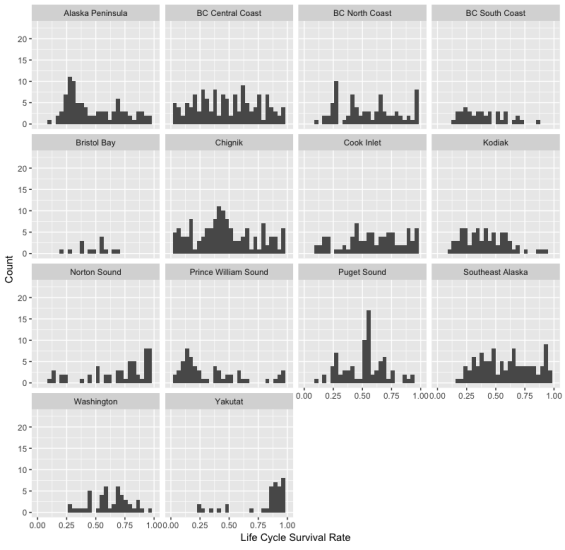
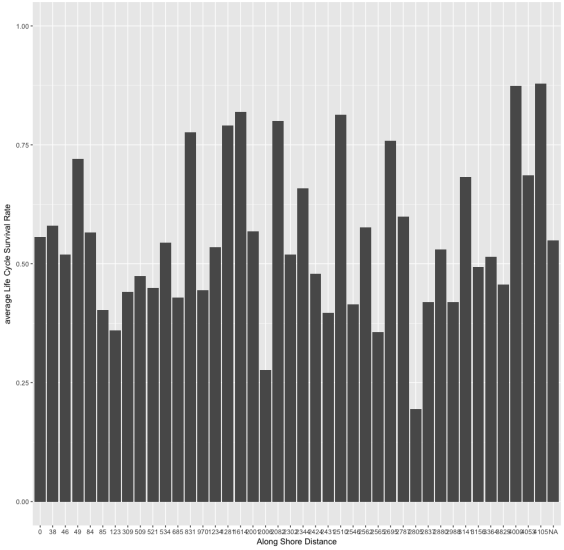


Figure 4: Life Cycle Survival Rate for Stock Breeding Broken Down by Region with Color for Stock Population

Figure 5: Along Shore Distance vs Life Cycle Survival Rate



5.2: Model 1 Assumptions, Interpretations, Settings and Trace Plots

Assumptions:

- Conditional independence of the spawn counts for a given stock brooding given the recruit counts for that stock brooding and life cycle survival probability for that stock brooding
- Prior independence in our uncertainty about the regional effects α_j

Interpretations:

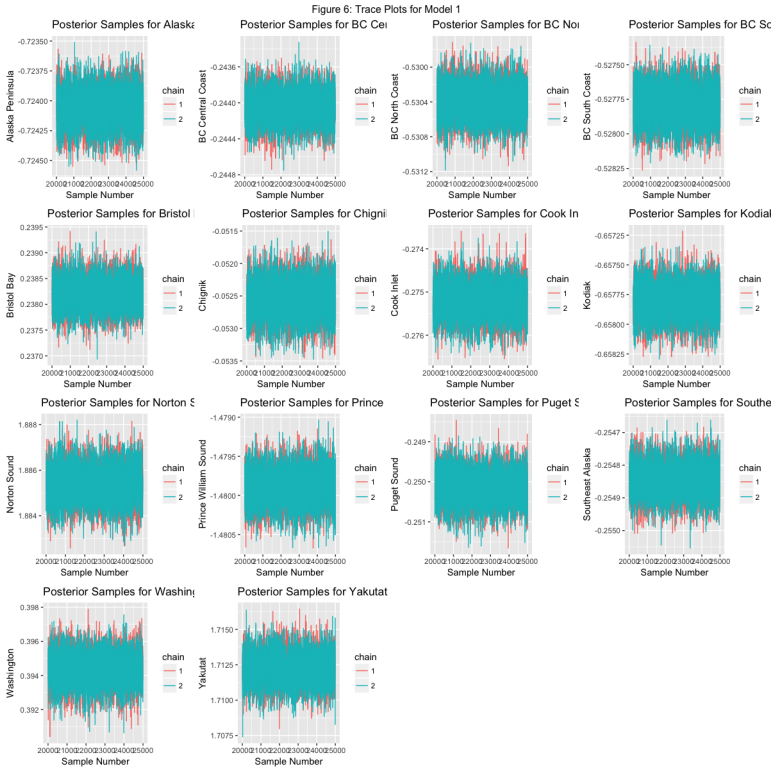
α_j: The effect of brooding in region j on the logit probability of survival (from recruitment to spawning) of a stock brooding in region j.

Algorithm Settings

All α_j were initialized to be zero. The following settings were used:

adaption steps: 10000
burn in steps: 10000
number of chains: 2
number of saved steps: 10000
number of thinning steps: 1

Trace Plots



5.3: Model 2 Assumptions, Interpretations, Settings and Trace Plots

Assumptions:

- Conditional independence of the spawn counts for a given stock brooding given the recruit counts for that stock brooding and life cycle survival probability for that stock brooding
- Conditional independence of the stock brooding random effects given τ^2

- Conditional independence of the regional effects given β_2 and τ_α^2

Interpretations:

β_0: There is no useful interpretation for this parameter, as the time parameter in the model represents the year from 1950, which is never 0.

β_1: Given a fixed region, 2*β_1 represents the effect of a 2 year increase in the brood year for a stock on the survival probability (from recruiting to spawning).

α_j: Adjusting for the year of breeding, this is the effect of brooding in region j on the logit probability of survival (from recruitment to spawning) for the breeding of a stock in region j.

γ_i: For a stock spawning in a given year, this represents a random effect that is used to account for overdispersion.

τ^2: The individual variation associated with the breeding of a stock in a given year.

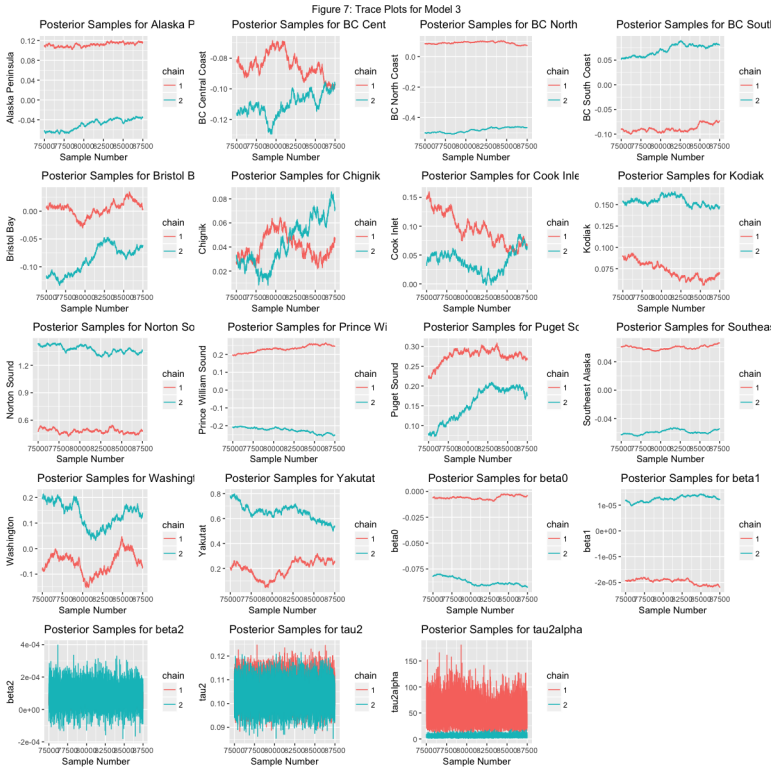
β_2: For a one unit increase in the median along shore distance of breeding sites within a region, there is a β_2 unit increase in the effect of region on the logit life cycle survival probability.

Algorithm Settings

All of the α_j were initialized to 0. All β parameters were initialized to zero. All τ parameters were initialized to 1.

adaption steps: 50000
burn in steps: 25000
number of chains: 2
number of saved steps: 25000
number of thinning steps: 1

Trace Plots



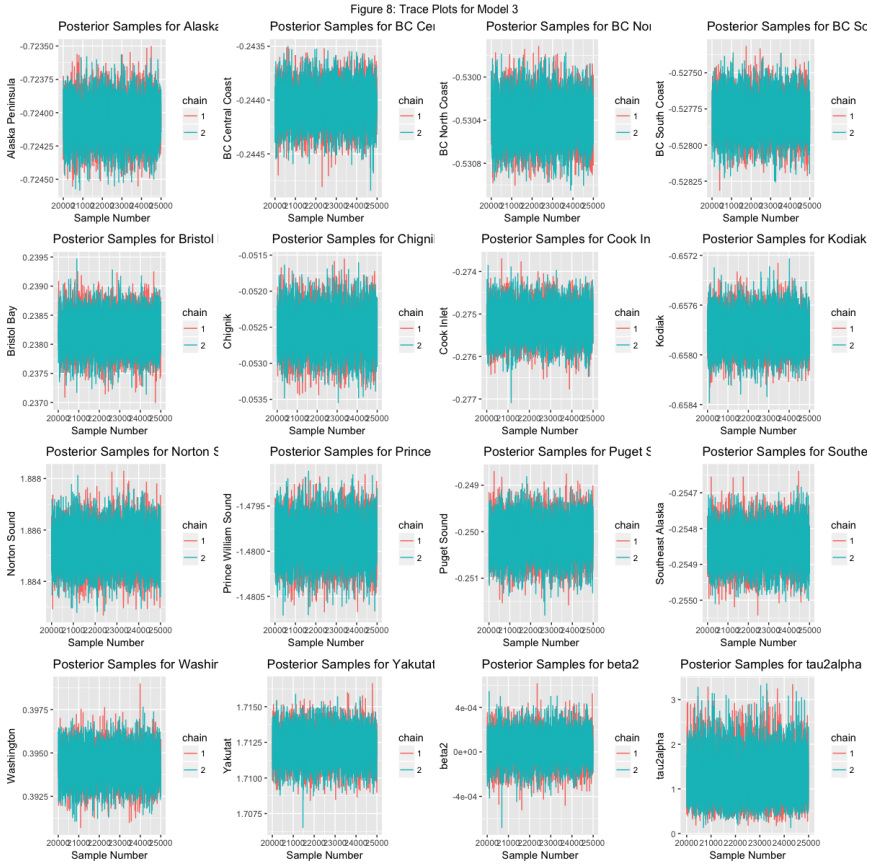
5.4: Model 3 Settings and Trace Plots

Algorithm Settings

τ_α^2 was set to 1. β_2 was set to 0. all α_j effects were set to 0.

adaption steps: 10000
burn in steps: 10000
number of chains: 2
number of saved steps: 10000
number of thinning steps: 1

Figure 8: Trace Plots for Model 3



5.5: Figures for Addressing Question 1

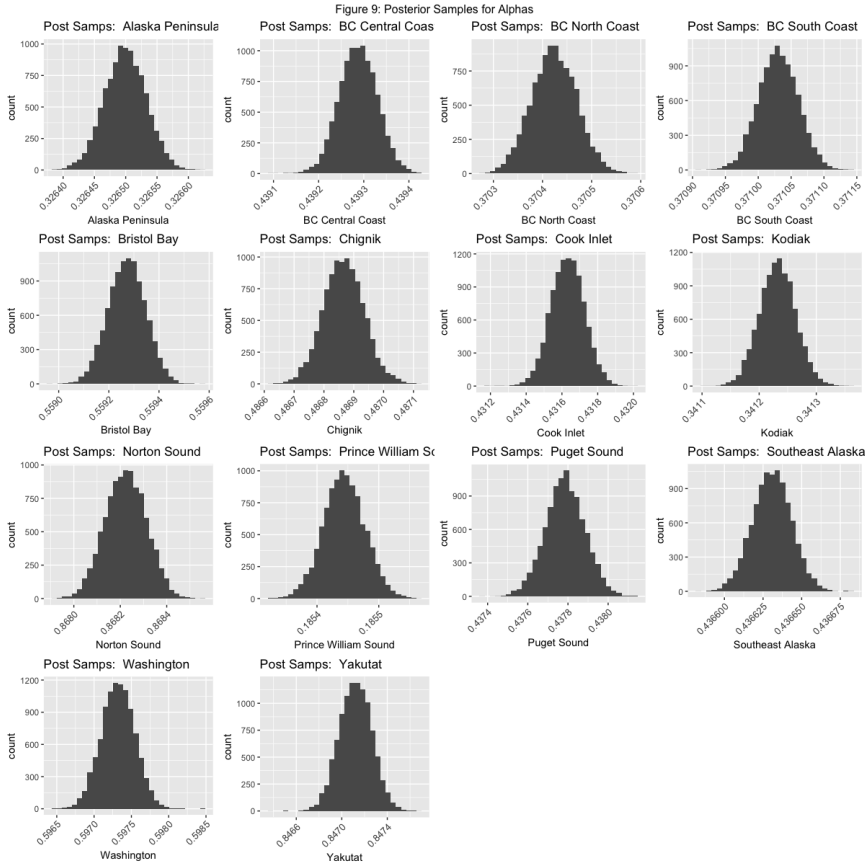
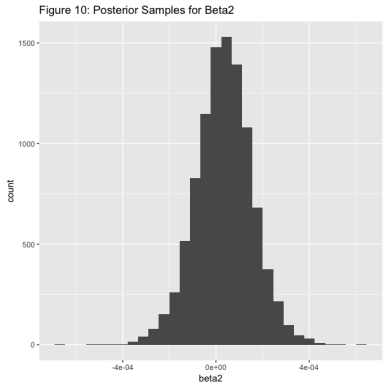


Table 1: 2.5th and 97.5th Posterior Quantiles of the Probability of Salmon Survival By Region

AK Pen	BC Cen	BC Nor	BC South	Bristol	Chignik	Cook In	Kodiak	Nort Snd	Prnc W Snd	Pug Snd	SE AK	Wash	Yakutat
0.32643	0.43921	0.37033	0.37097	0.55913	0.48673	0.43145	0.34117	0.86805	0.18538	0.43759	0.43661	0.59687	0.84684
0.32656	0.43937	0.37051	0.37109	0.55941	0.48700	0.43182	0.34129	0.86839	0.18551	0.43798	0.43666	0.59776	0.84740

5.5: Figures for Addressing Question 2



6: References

Brigitte Dorner, Randall M. Peterman, and Zhenming Su. Evaluation of performance ofalternative management models of pacific salmon (oncorhynchus spp.) in the presence ofclimatic change and outcome uncertainty using monte carlo simulations.Canadian Journal ofFisheries and Aquatic Sciences, 66(12):2199–2221, 2009.

NOAA Office of Science and Technology (2015). Fisheries of the United States. Retrieved from https://www.st.nmfs.noaa.gov/commercial-fisheries/fus/fus15/index

Oksana Chkrebtii and Jiguo Cao. Modeling spatiotemporal trends in the productivity of NorthPacific salmon.Environmetrics, 24(1):31–40, 2013. ISSN 1099-095X. doi: 10.1002/env.2183.URLhttp://dx.doi.org/10.1002/env.2183.