Comparing Stickleback Fish Over Time and Space

Ellen Cho, Isabel Heard, & Basia Sychla

12/11/2023

Introduction

For the class project, we worked with Dr. Yoel Stuart, an evolutionary biologist, to explore the phenotypic traits occurring in the stickleback fish species. In particular, we were interested in whether there is a difference in the traits found near salty bodies of water versus in fresher water as well as whether the same patterns can be found years later or in different locations.

We were given a total of five main data sets. Two were for Bonsall Creek, one for 2006 and one for 2020. Three data sets were from Campbell River and Courtenay River. The study explored whether there was a significant change in clines over time (2006 vs. 2020) and space (Bonsall Creek vs. Campbell River vs. Courtenay River). Salinity levels (concentrations of salt in the water) were also measured or calculated for each location. Additionally, CFit7, a software used in the field to calculate clines, was also introduced to our group.

Methods

Software

Since the 2006 data was from a published paper (Vines et al.), our initial goal was to recreate these results. We attempted to create input files for CFit7 as was done in the 2006 study, but the results were either not the same as the published results or the input file was not being read successfully. A lot of time was spent trying to get CFit7 to work properly, but given the limited time we had for this project, we decided it was not the way to move forward. Therefore, we continued our analysis by using R.

Data Compilation & Cleaning

The data sets were given to us in Excel spreadsheets. A lot of clarifications were needed in regards to the traits and their equivalents across data sets. Some of the traits were measured in pixels and needed to be converted into millimeters using a conversion column. It was also necessary to calculate distances from GPS coordinates for Courtenay and Campbell rivers. For Campbell river, we excluded data for locations B1 and B2 as they refer to pockets of water near the river. Right away, we noticed notable outliers. After consulting with Dr. Stuart, he informed us that he usually removed outliers that were three standard deviations away from the mean. Since there was some notable skewness in the data, we decided to set the criteria to three standard deviations away from the median.

Furthermore, we were given R code to calculate salinity from the measurements that were taken at Bonsall Creek, and a spreadsheet with salinity levels for the two rivers. However, different salt levels were recorded at the surface and bottom water levels at any given point. In general, the bottom of the river or creek contained more salt than the surface, so there was some confusion as to which measurements to use. Unfortunately, the question of how to use salinity

levels will have to be resolved in the future. For the purposes of our study, we agreed with our collaborator that we should use distances from the estuary as the independent variable. Once this was resolved, we narrowed down our focus to study the five traits that were highlighted in the Vines study and which happened to be available for each of the locations and time periods.

The five traits we studied were: pelvic spine length, second dorsal spine length, dorsal fin length, caudal peduncle depth, and pectoral fin length. Based on the results published in the Vines paper and previous studies, it was hypothesized that each of the five traits tend to decrease in size the farther a fish lives from the estuary. Also, the water close to the estuary is expected to be an unstable environment. Therefore, it is theorized that in those areas, longer fins are needed for stability.

Model Selection and Comparison

One of our initial goals was to recreate the published data in the Vines paper and to make comparisons. However, near the end of the semester, we realized that we were using the full data set while the authors of the Vines paper were using about a half of the available observations. It was not clear to us what criteria was used to exclude these observations. We concluded that this was most likely the reason our results were differing greatly from that of the paper.

To determine the best model to fit the data, we consulted with Dr. Matthews. Dr. Matthews emphasized that a cline was simply a logistic model, and that while the equation of the model used in CFit7 was slightly different, it had basically the same foundation as a logistic model (asym/(1+exp((xmid-input)/scal))). Therefore, we decided to move forward with an NLS (Nonlinear Least Squares) model. Since we did not know the starting values to begin the iterations to fit this model, he advised us to use the SSLogis function (self-starting NLS logistic regression model) as follows:

```
Model <- nls(residuals ~ SSlogis(distance, asym, xmid, scal), data = trait_data)
```

The Logistic Model (SSlogis) graph in the Appendix shows how each of the three parameters (asym, xmid, and scal) can be visualized. We attempted to fit this model for each of the traits for each creek or river. However, it appears that the model cannot be fit if the data does not resemble a logistic regression. If R was successfully used to fit such a model, the data points, a red model line, and center point have been graphed and included in the Appendix. Additionally, the model parameter estimations may be found in the Appendix. Otherwise, if a model could not be fit, the data points along with a spline in blue has been graphed instead.

Finally, we consulted Dr. O'Brien to determine how to compare the parameters obtained from the NLS models (asym, xmid and scal). He suggested using an ANOVA where we could compare full models (where all the parameters of the NLS models were different) and reduced models (where one of the parameters was shared between the NLS models). Dummy variables were set to indicate which group the data came from. The full model contained separate parameters for asym, xmid, and scal (i.e. th2A = xmid for 2006 data, th2B = xmid for 2020 data, etc.). For the reduced models, either xmid (th2) or scal (th3) was set to be shared by the two

different data sets (LL3modB and LL3modC in code below). If the ANOVA showed that the reduced model was preferred (giving us the F-statistic and p-values for the comparisons), we could say that the model where the center (or slope) was shared between the two datasets was a better fit, indicating that there was no significant difference for that parameter between the two data sets.

```
LL3modA <- function(x,th1A,th1B,th2A,th2B,th3A,th3B) {
  th1 <- th1A*data$dumA+th1B*data$dumB
  th2 <- th2A*data$dumA+th2B*data$dumB
  th3 <- th3A*data$dumA+th3B*data$dumB
  th1/(1+exp((th2-x)/th3))
}

LL3modB <- function(x,th1A,th1B,th2,th3A, th3B) {
  th1 <- th1A*data$dumA+th1B*data$dumB
  th3 <- th3A*data$dumA+th3B*data$dumB
  th1/(1+exp((th2-x)/th3))
}

LL3modC <- function(x,th1A,th1B,th2A,th2B,th3) {
  th1 <- th1A*data$dumA+th1B*data$dumB
  th2 <- th2A*data$dumA+th2B*data$dumB
  th1/(1+exp((th2-x)/th3))
}
```

Results

NLS Models and ANOVA

One of the traits we have explored was the depth of the caudal peduncle. For the 2006 data, we can see that as we get to the middle of the creek, the residuals increase by a little, but not much. In the 2020 data, the spline is fairly straight, we think this is because in the 2020 data, it mostly consists of fresh water. This could tell us that we might not see a whole lot of changes in the trait size. Unfortunately, for this trait, we were unable to run our NLS model. Therefore, all graphs for caudal peduncle were fitted with a spline (seen in blue) and included in the Appendix.

Next, we looked at the pelvic spine length. For the 2006 data, it appears that as the stickleback were caught further away from the estuary, the smaller their pelvic spines. We were able to fit a logistic regression model to the 2006 data, and the graph has been included in the Appendix. We found that we have a midpoint (center) of 2.6, an asymptote of -0.48, and a scale parameter (representative of the inverse of the slope at midpoint) of 0.08. For the 2020 data, we were not able to get the NLS model to converge, so only a spline could be used as a model of the data points. Even so, it does appear that the general pattern of the trait stayed fairly consistent over time.

For the second dorsal spine length, after visualizing the data, we found that it looked very similar for both time periods. N LS models were successfully fitted onto the data sets, and the

centers were 2.15755 and 1.9779 (for 2006 and 2020 respectively). The scale parameters were -0.12199 and -0.1440. After comparing the parameters through an ANOVA, we found that with a p-value of 0.42 and 0.7796, the centers and the slopes were not significantly different. Therefore, the centers and the slopes of this trait did not change from 2006 to 2020. Dorsal fin length gave us some trouble when it came to fitting an NLS model for the 2020 data. Therefore, the two data sets could not be compared through an ANOVA. The center for the 2006 data showed a center at 2.213 and a scale parameter of -0.0886.

Pectoral fin length was fitted to an NLS model for both years. 2006 had a center at 3.5263 and a scale parameter at 0.1525. For the 2020 data, residuals below -3 had to be removed in order to fit the model. Both the unfitted and fitted data are available in the Appendix (spline fitting in blue, NLS fitting in red with residuals removed). The 2020 data fitted to the NLS model had a center at 2.219 and a scale parameter at -0.10417. We were not very confident with models for this trait because the scale parameter came up positive in 2006 but negative in 2020 despite the way the data looked when plotted. For the ANOVA, as expected, the slope comparison did not converge, but the center comparison showed that there was a significant difference between the centers of the 2006 data and the 2020 data. This means that the centers for pectoral fin length are significantly different between 2006 and 2020.

Finally, the analysis of Campbell and Courtenay river data provided a different perspective. None of the traits measured for either of the two rivers showed a typical cline pattern. In fact, the spline plots (included in the Appendix) each showed nearly horizontal lines. It was, therefore, not possible to fit a logistic regression model for any of the traits from these rivers.

Conclusion

Throughout the course of working on this project, we learned a lot about working in academia and consulting many different professionals to obtain the best results possible. Being that evolutionary biology was very new to all of us, our first few meetings were all about understanding the stickleback fish and the background information that came along with this project. We had a large collaborator and expert pool to gather information from that helped us with our analysis. This also meant that it would take some time for our questions to get answered. We also learned that it is very difficult to recreate work from a published paper.

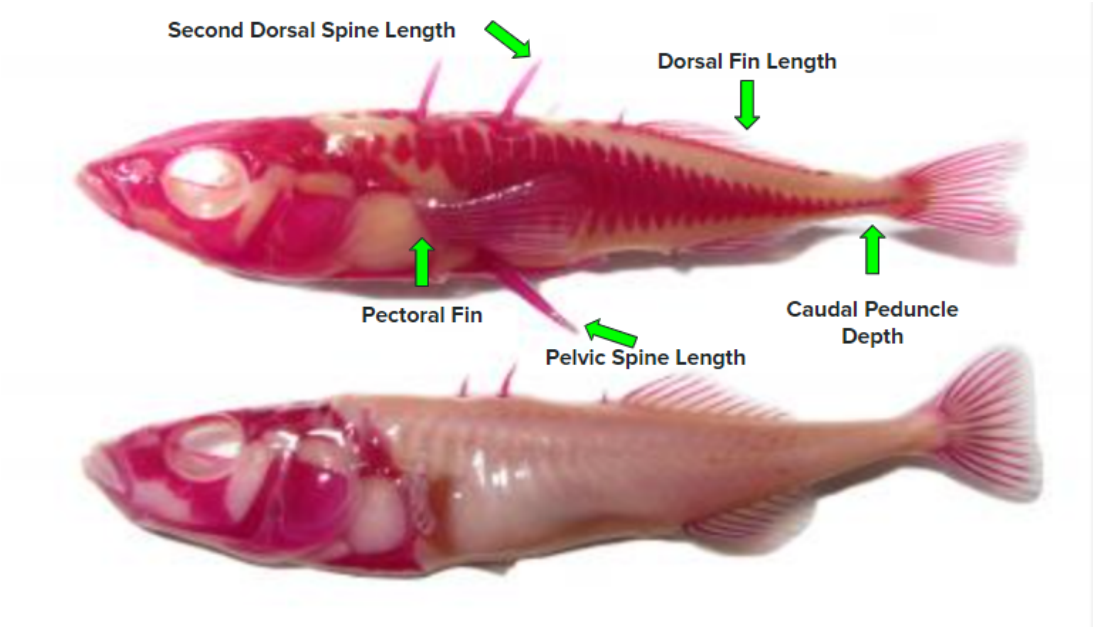
Overall, it is difficult for us to make concrete conclusions about the analysis we have done. Our datasets exhibited numerous unexplainable distinctions, rendering the task of drawing meaningful comparisons quite challenging. The 2020 data for Bonsall Creek was mostly freshwater, which calls to question the effect that salinity has on the change in traits we see. Further studies would need to be done to flesh out the cause and effect between traits and environmental variables. Also, comparing Bonsall (2020) to Courtenay and Campbell (2021) could be difficult since the cause of the change in traits could be difficult to pinpoint with all the various variables.

One possible cause for why many of our NLS models did not converge was the shape of the data. Dr. O'Brien informed us that sometimes it is possible to force a certain model onto the data, but if it doesn't fit, it will not fit. In our case, we believe that one of the reasons why our initial work with CFit7 was not fitting or reading properly was due to the shape of the data. Once we found out that the Vines paper only analyzed about half of the given data, we realized that this may have played a large role in why our input files would not give us the same results. If we had more time for this project, we'd definitely look into the criteria that was used to remove the data and try again (using CFit as well as R). Additionally, we'd likely take a closer look at salinity levels and the other traits that weren't necessarily provided for each of the locations or years we covered.

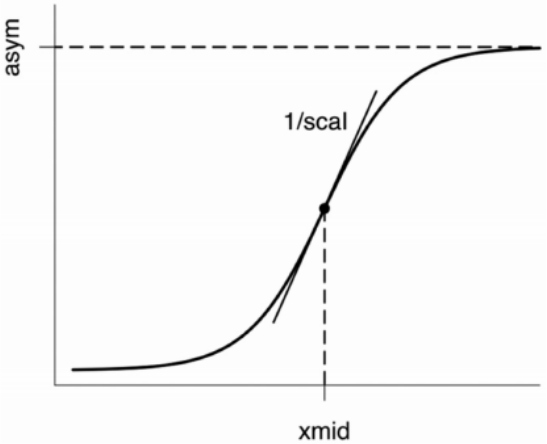
One suggestion we have for future work on this topic is to reorganize how data is recorded. A more straightforward way of data storage may save a lot of time during any future analyses. Also, we believe the evolutionary biology field would greatly benefit from an updated version of CFit7. While the program itself has the potential to provide a lot of meaningful calculations, it would be very helpful if the process of inputting data was more straightforward and user-friendly. Furthermore, we believe it would be a good idea to closely study the stickleback species by finding a way to track individual fish throughout their lifespans. While the species already offers a wealth of knowledge about evolution, specific phenotypic or genetic traits might be better understood if we knew how far a fish travels in its lifespan, and whether the mating and egg development occurs exclusively in the estuary.

Appendix

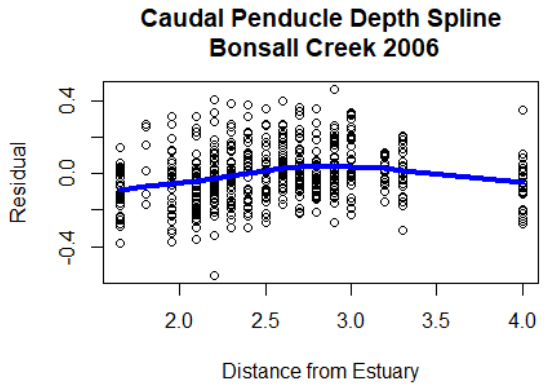
Traits that were studied in this project

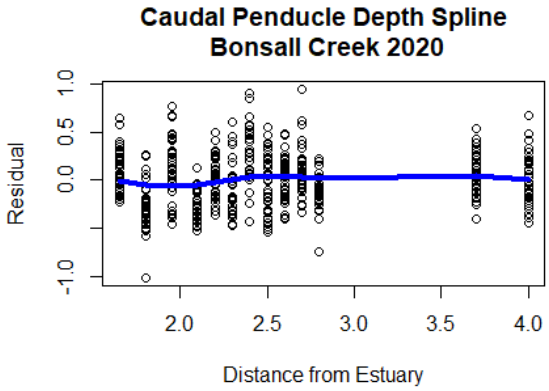


Logistic model (SSLogis)

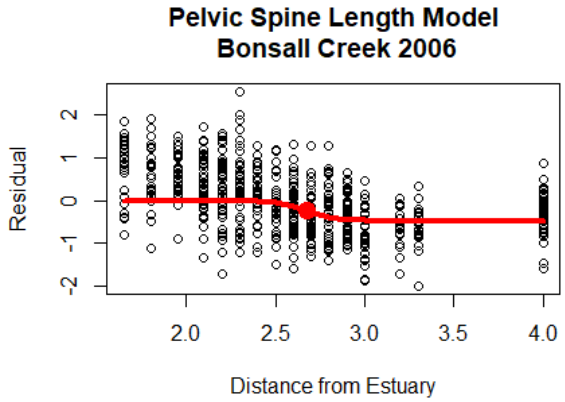


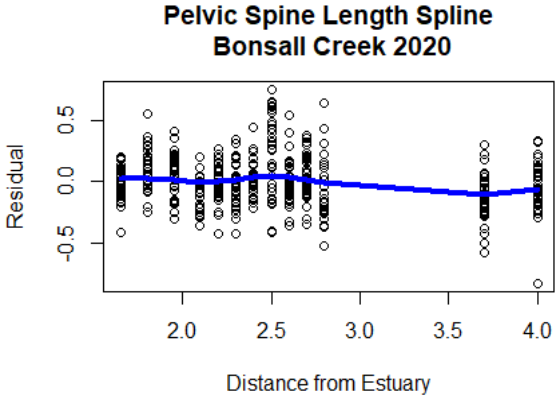
Caudal Peduncle 2006 vs 2020 (fit withs splines in blue)





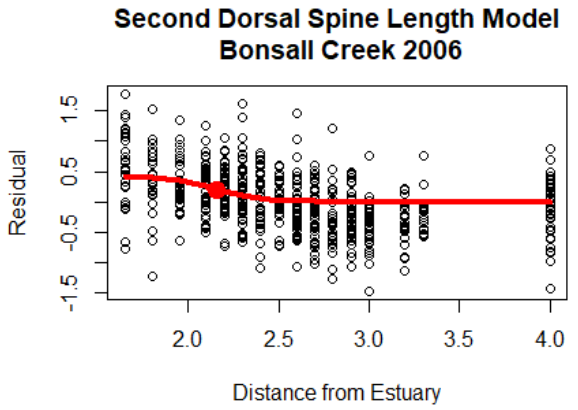
Pelvic Spine 2006 vs 2020 (NLS fit in red, splines fit in blue)

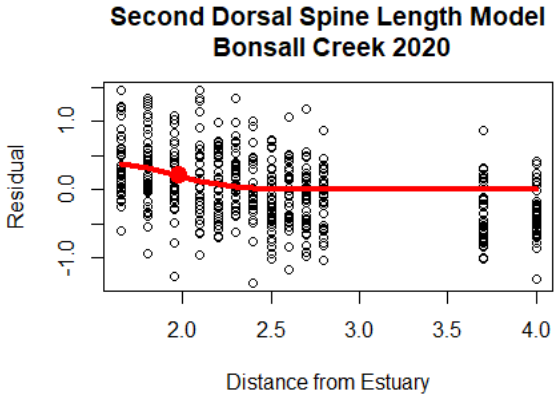





```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym -0.47574    0.06177  -7.701 5.71e-14 ***
xmid  2.67697    0.05623  47.606  < 2e-16 ***
scal  0.08114    0.05084   1.596    0.111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6984 on 590 degrees of freedom

Number of iterations to convergence: 29
Achieved convergence tolerance: 9.138e-06
```

Second Dorsal Spine 2006 vs 2020 (NLS fit in red)

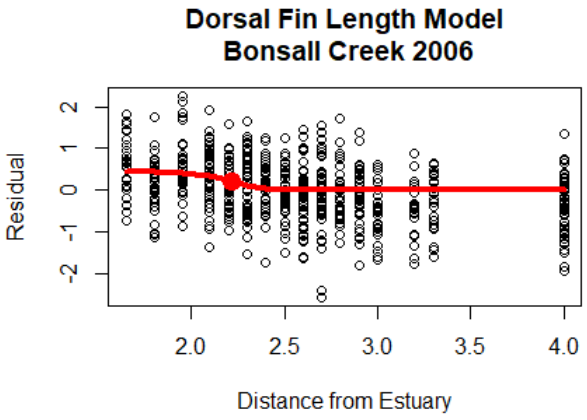


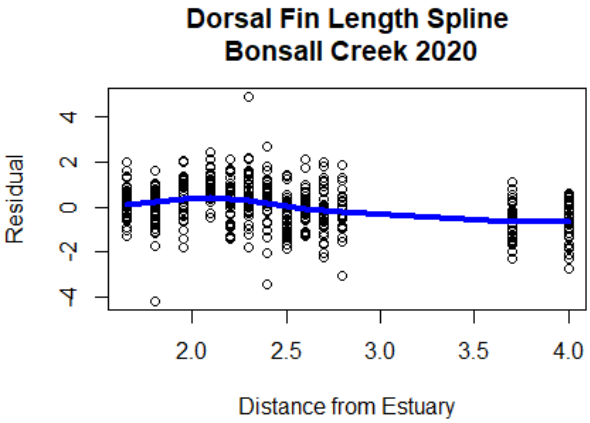



```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym  0.42892    0.08123   5.280 1.83e-07 ***
xmid  2.15755    0.08078  26.707  < 2e-16 ***
scal -0.12199    0.06310  -1.933   0.0537 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4764 on 578 degrees of freedom

Number of iterations to convergence: 25
Achieved convergence tolerance: 8.707e-06
```

```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym   0.4097     0.1632   2.510   0.0124 *
xmid   1.9779     0.1794  11.023   <2e-16 ***
scal  -0.1440     0.1026  -1.404   0.1611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5068 on 466 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 6.393e-06
```

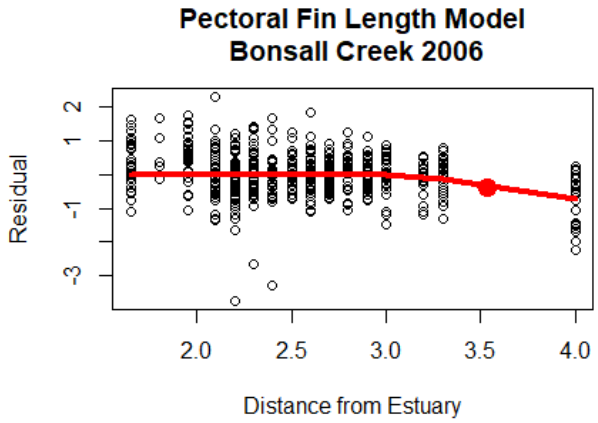
Dorsal Fin 2006 vs 2020 (NLS fit in red, spline fit in blue)

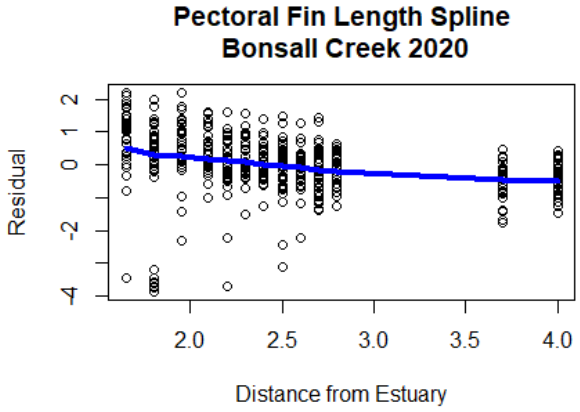


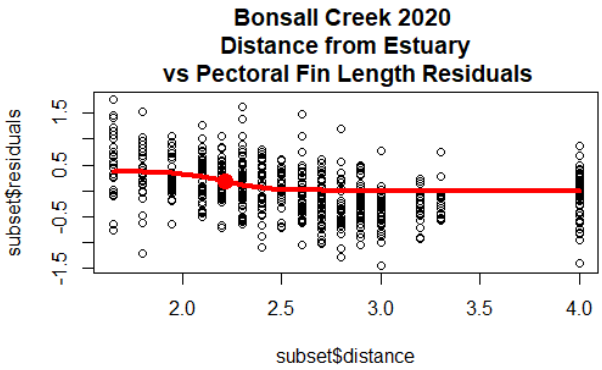



```
Parameters:
     Estimate Std. Error t value            Pr(>|t|)
Asym  0.43620    0.08668   5.032          0.000000644 ***
xmid  2.21346    0.07450  29.712 < 0.0000000000000002 ***
scal -0.08861    0.06417  -1.381                0.168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.712 on 594 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 0.000003608
```

Pectoral Fin 2006 vs 2020 (NLS fit in red)



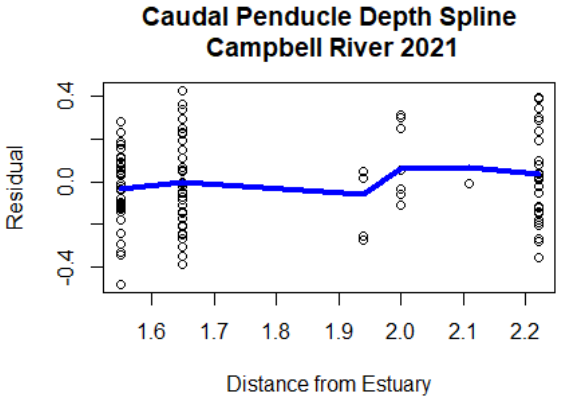


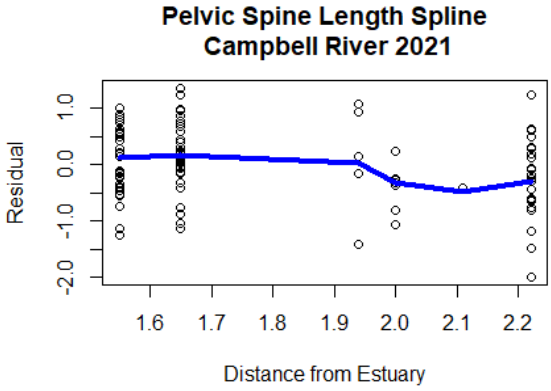


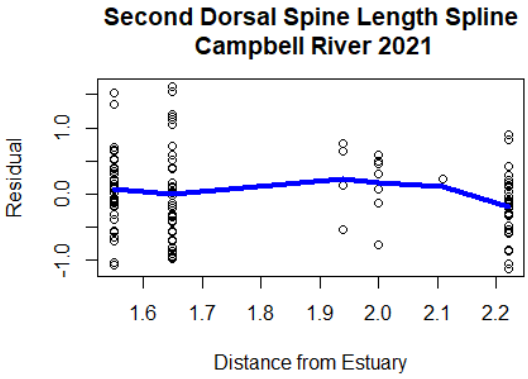

```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym  -0.7787     0.3108  -2.505   0.0125 *
xmid   3.5263     0.5275   6.685 5.93e-11 ***
scal   0.1525     0.2489   0.613   0.5405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6074 on 525 degrees of freedom
```

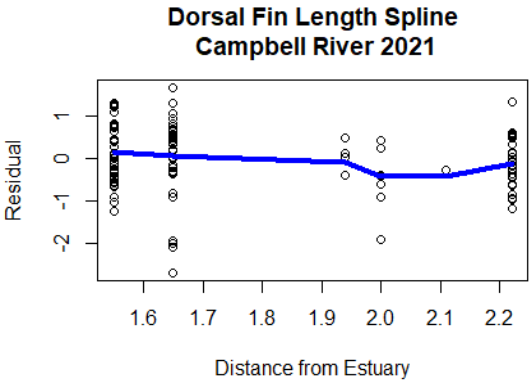
```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym  0.37336    0.06495   5.748 1.46e-08 ***
xmid  2.21917    0.06979  31.799  < 2e-16 ***
scal -0.10417    0.05801  -1.796   0.0731 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4704 on 574 degrees of freedom
```

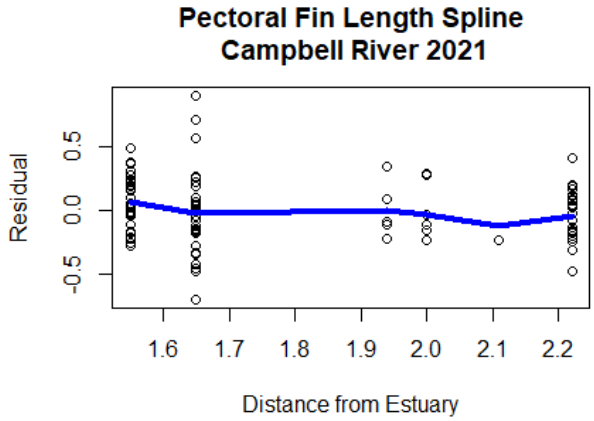
Campbell River











Courtenay River