

Analysis of Rare Fish Data (November 23, 2012)

by Tim Robinson(University of Wyoming)

For the Tuxedo Darter and the Spotfin Chub, we will consider two potential questions of interest:

1. For a given species (i.e. Tuxedo Darter and Spotfin Chub), are there any locations which show a trend over the existing years of data? If so, what is the magnitude of the trend? These questions are ‘location-specific’ questions since they allude to specific sampling locations.
2. Overall, for a given species (i.e. Tuxedo Darter and Spotfin Chub), is there a trend over the existing years of data? If so, what is the magnitude of the trend? These questions are ‘study area averaged’ questions since they allude to the study area as a whole.

Sampling Unit-Specific Analyses Tuxedo Darter

Let’s refer to the first question above as a ‘location-specific’ question since we will look at separate analyses of each individual sampling location for each species. To address this question, we can simply fit the following linear model to each of the 5 sampling locations for the Tuxedo Darter:

$$CPUE_{ji} = \beta_{0i} + \beta_{1i}Year_{ji} + \epsilon_{ji} \quad (1)$$

where $CPUE_{ji}$ denotes the observed (Catch per Unit Effort)CPUE for the j^{th} year at the i^{th} sampling location ($i = 1, 2, \dots, 5$), β_{1i} denotes the annual trend (i.e. average annual change) for the i^{th} sampling location, and $Year_{ji}$ denotes Year j for the i^{th} sampling location. The above regression model along with the mixed model that we will run later assumes that the model errors (i.e. the ϵ_i) have constant variance. We will utilize a Cleveland plot [see Cleveland (1993)] to visually assess the validity of this assumption. If the variance is non-constant, we can utilize a transformation of the response to address the heterogeneity. Using R (coded provided upon request) the Cleveland plots for the untransformed (left panel) and the log-transformed data (right panel) are provided below. Note that the spread within each of the clusters is similar across the two panels, thus suggesting that a log-transformation of the response is unnecessary. Further residual analyses were conducted and these analyses further verified that no transformation was needed for the modeling of CPUE.

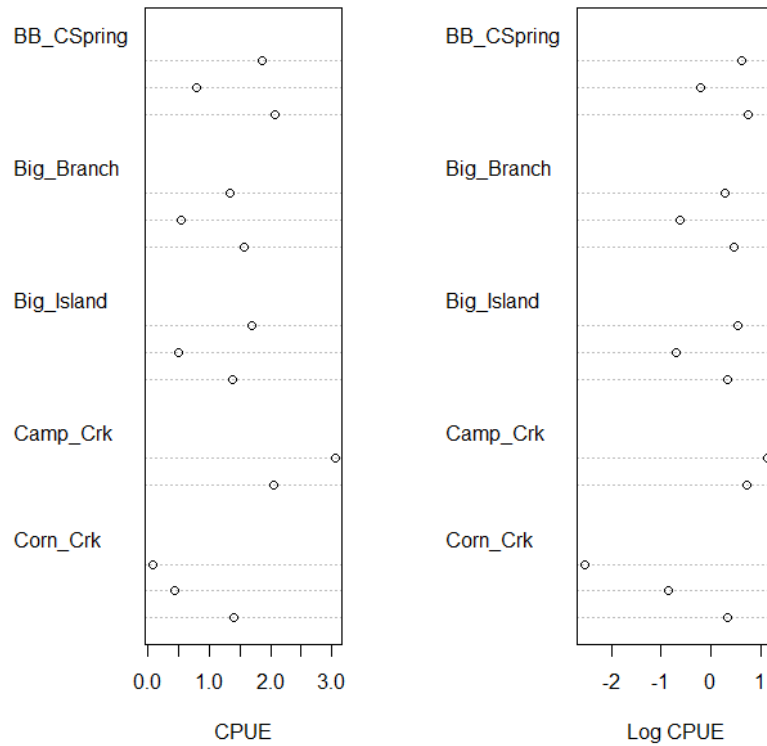


Figure 1. Cleveland plots of Tuxedo Darter CPUE values (left-panel) and log-transformed CPUE values. Since the dispersion within each set of clustered observations is similar across the two panels, a log-transformation of the response was deemed unnecessary.

Fitting the model in (1) to the five sampling locations we observe the 95% confidence intervals in Figure 2 when considering the intercepts (i.e. the β_{0i}) and slopes (i.e. the β_{1i}) for the fitted trends in Tuxedo Darter CPUE over the study years. The red reference line in the right-panel of Figure 2 denotes ‘no average annual change’ across the existing years of data. Note that the right-panel confidence intervals associated with each sampling location covers the ‘no average annual change’ value (i.e. slope for Years = 0) and thus there has been no significant trend in the averaged CPUE over the existing years of data for any of the sampling locations. Figure 3 displays the observed CPUE values for each sampling location along with the linear regression fits the CPUE values at each of the sampling locations. It should be noted that only Parch Corn Creek was observed to show a negative trend but this trend was not a statistically significant trend ($p = 0.347$). An interesting feature of the fits to each sampling location is that the fitted 2009 CPUE values are all systematically larger than the observed 2009 CPUE numbers. Removing the 2009 CPUE values would have resulted in slight negative trends at the Big Branch and Big Branch/Cold Spring locations instead of the currently observed positive trends at each of these locations.

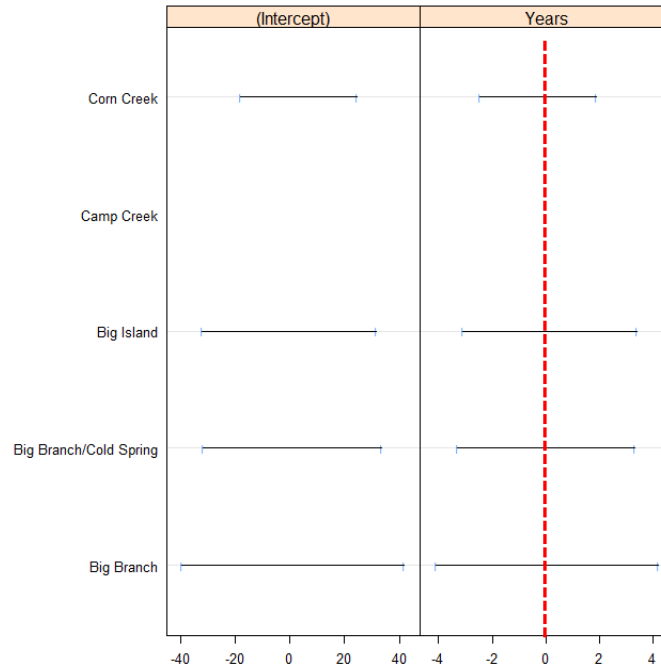


Figure 2. The 95% confidence limits on the intercepts (left-panel) and slopes (right-panel) for the fit of the linear regression model given in (2) for each of the five sampling locations. Red reference line indicates a value of ‘no average annual change’.

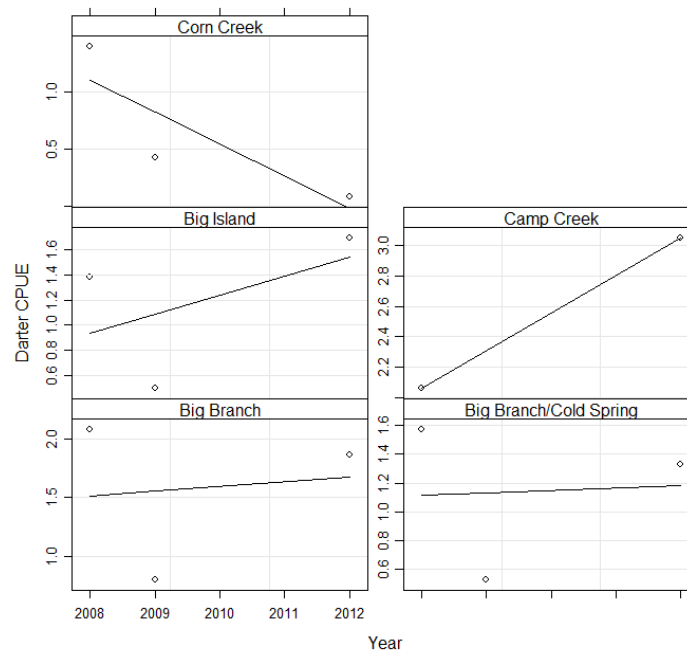


Figure 3. Tuxedo darter CPUE's with the fitted linear regression model in (1) for each of the five study plots. While fits appear to suggest slight trends, when comparing the observed trend to the calculated standard errors, trends are not statistically significant. The lowest p-value is associated with Parch Corn Creek(Corn Creek) $p=0.347$.

Study Area Averaged Analyses (Tuxedo Darter)

This section will address the second set of questions and we will refer to these set of questions as ‘study area averaged’ questions since we wish to describe what is happening in general across all five sampling locations. Before explaining the model in this section, it is important to note that for the Tuxedo Darter, since no significant trend was observed for any of the five study locations, we will not observe a significant trend for the average trend across the five locations. To address the second set of questions, we can simply fit the following linear model to the pooled data (i.e. pool the data across all five study plots):

$$Count_{ji} = \beta_0 + \beta_1 Year_{ji} + \epsilon_{ji}. \quad (2)$$

In comparing the model in (2) to that in (1), note that there is no ‘*i*’ subscript on the intercept or on the slope coefficient, suggesting that a common ‘study-area wide’ intercept and slope are fit to the entire set of data. The model in (2) does not allow for any differences in the sampling locations since we are averaging over each sampling location. The model in (2) also does not account for the fact that repeated observations on each sampling location (i.e. repeated over time) are correlated with one another. To determine if CPUE values differ depending upon the sampling location, we can examine boxplots of the standardized residuals from the fit of the model in (2) to see if the boxplots are systematically above and below a reference line of zero according to the sampling location. If any of the boxplots are entirely above or below zero, we need to fit a model that allows for differences among the sampling locations. Figure 3 shows the boxplots of the standardized residuals from the fit of the model in (2) according to sampling location. The standardized residuals from Station Camp Creek are all above zero and most of the residuals associated with Parch Corn Creek are below zero, thus suggesting the need to account for differences in the sampling locations. The following *linear mixed model* will be fit to the data:

$$CPUE_{ji} = \beta_0 + \beta_1 Year_{ji} + \delta_{0i} + \epsilon_{ji}. \quad (4)$$

The model in (3) differs from the one in (2) by the random intercept component δ_{0i} . This component can be interpreted as the random deviations across sampling locations in terms of overall location abundances. So, using CPUE as a proxy for ‘abundance’, we observe a larger abundance of Tuxedo Darters at the Station Camp Creek location than the other sampling locations.

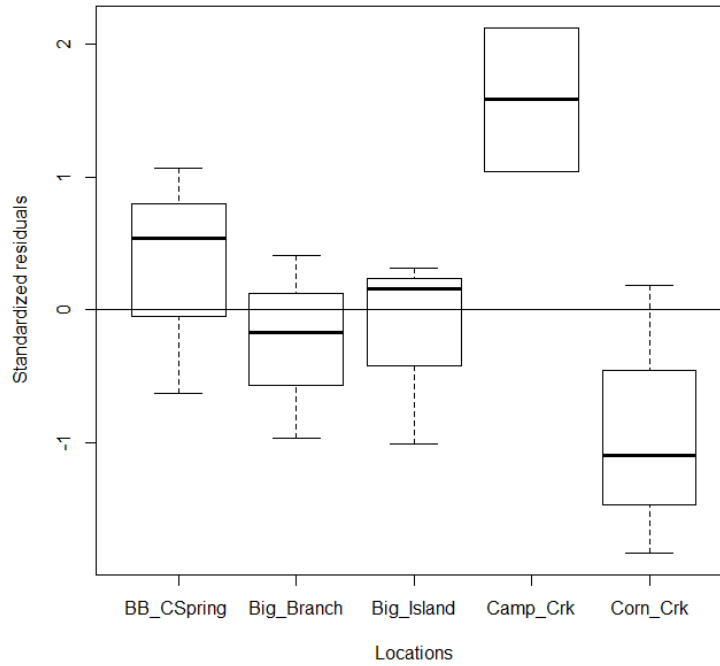


Figure 4. Boxplots of the standardized residuals from the fit of the model in (2). Note that for Station Camp Creek (Camp_Crk) the standardized residuals are systematically larger than zero and for Parch Corn Creek (Corn_Crk) the standardized residuals are predominantly below the zero reference line. This suggests that we need to account for differences in the sampling locations when fitting the study-area averaged model.

Fitting the model in (3) using the ‘lme’ function in R, we observe results in Table 1. The results in Table 1 are broken down into two parts: 1. Random effects and 2. Fixed effects. The random effects portion indicates the sources of variability observed in the existing data. The total variation in CPUE values is partitioned into location to location variance (i.e. $0.5154^2=0.2656$) and ‘within location variability’ (i.e. $0.6970^2=0.4858$). The total variation in CPUE is then $0.2656+0.4858 = 0.7514$ and the ‘within location variability’ is estimated to account for $\left(\frac{0.4858}{0.7514}\right)100\% = 64.65\%$ of the total variation in CPUE values. It is important to note that the model in (3) assumes equal trends across all five sites and, as observed in Figure 3, Parch Corn Creek appears to have a slight decline in CPUE values while the other sites appear to have slight increases in CPUE values. Most of the variation that is currently attributed to ‘within site variation’ is due to the observed 2009 CPUE values. As mentioned earlier, the predicted CPUE values for 2009 are larger than the observed CPUE values at all sites. While the existing data

suggests most of the variation can be attributed to ‘within site variation’, it is quite possible that, in actuality, most of the variation is due to the variation across sites. In most ecological settings, the spatial variation (i.e. variation across sites) is larger than the ‘within site’ variation. In attempting to fit different slopes for each of the sites utilizing the linear mixed model, we were not able to obtain convergence in the estimation of the model parameters. This is not surprising given the limited amount of data and the very slight trends that are currently being observed.

The fixed effects portion indicates the estimated intercept and slope from the fit of the model in (3). Note that the estimated slope is 0.0397, suggesting a slight positive trend in the overall average CPUE values across sites. The associated p-value of 0.7165 indicates that this observed trend is not statistically significant. Thus, as expected (when considering that no significant trend was observed for any given sampling location), no significant overall trend across the study years exists for the Tuxedo CPUE values.

Table 1. Mixed model output from the fit of the model in (3) to the Tuxedo Darter data

Random effects:

Formula: ~1 | Location
 (Intercept) Residual
 StdDev: 0.5154 0.6907

Fixed effects: CPUE ~ Year

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.9975	1.0691	8	0.9330	0.3781
Year	0.0397	0.1056	8	0.3763	0.7165

Statistical Power for Tuxedo Darter

It is important to keep in mind that just because we did not observe a statistically significant trend across the sites in terms of Tuxedo Darter CPUE, this finding does not necessarily mean that the Tuxedo Darter abundance is not declining. With only four locations with which to estimate trend, we likely have a very low probability of detecting a trend if a trend actually exists. The ability to detect an actual trend is commonly referred to as the statistical power of a monitoring study. Generally speaking, for the types of models used to estimate trends in CPUE,

an examination of statistical power is conducted after a pilot set of data has been collected. The information from this pilot data is used to inform a simulation study. In the simulation study, site CPUE values are populated utilizing the ‘within site’ CPUE probability distributions. The amount of dispersion in this distribution is generally estimated to be the residual variance. For the Tuxedo Darter data, the residual variance is currently estimated to account for nearly 65% of the total CPUE variation. As was discussed earlier, it is quite possible that this value has a high amount of artificial inflation due to the observed 2009 values. While a power analysis could be simulated at this time, it is my opinion that we would be better served by having another year of data with which to better ascertain confidence in the 2009 observed CPUE values. Conducting a simulation at this point would likely result in highly conservative estimates of the power associated with the current monitoring plan.

Sampling Unit-Specific Analyses Spotfin Chub

To address the location-specific questions, we can simply fit the following linear model to each of the 9 sampling locations for the Spotfin Chub:

$$CPUE_{ji} = \beta_{0i} + \beta_{1i}Year_{ji} + \epsilon_{ji} \quad (5)$$

where $CPUE_{ji}$ denotes the observed CPUE for the j^{th} year at the i^{th} sampling location ($i = 1, 2, \dots, 9$), β_{1i} denotes the annual trend (i.e. average annual change) for the i^{th} sampling location, and $Year_{ji}$ denotes Year j for the i^{th} sampling location. We again utilize the Cleveland plot [see Cleveland (1993)] to visually assess the validity of this assumption. If the variance is non-constant, we can utilize a transformation of the response to address the heterogeneity. Using R (coded provided upon request) the Cleveland plots for the untransformed (left panel) and the log-transformed data (right panel) are provided below. Note that the spread within each of the clusters is much more similar for the clusters in the right-hand panel involving the log-transformed response. Further residual analyses were conducted and these analyses also verified that the log-transformation was necessary for fitting the Spotfin Chub CPUE values.

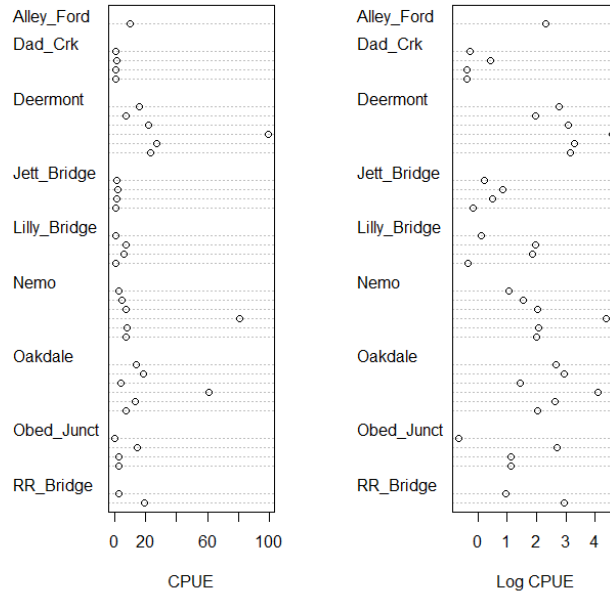


Figure 5. Cleveland plots of Spotfin Chub CPUE values (left-panel) and log-transformed CPUE values. Since the dispersion within each set of clustered observations is far more similar across the log-transformed clusters, we prefer the log-transformed CPUE values for modeling.

The model for the log-transformed CPUE values is given by:

$$\ln(CPUE_{ji}) = \beta_{0i} + \beta_{1i}Year_{ji} + \epsilon_{ji} \quad (6)$$

where $\ln(CPUE_{ji})$ denotes the natural log of the observed Spotfin CPUE for the j^{th} year at the i^{th} sampling location ($i = 1, 2, \dots, 9$), β_{1i} denotes the annual trend (i.e. average annual change) for the i^{th} sampling location, and $Year_{ji}$ denotes Year j for the i^{th} sampling location. Fitting the model in (6) to the nine sampling locations we observe the 95% confidence intervals in Figure 6 when considering the intercepts (i.e. the β_{0i}) and slopes (i.e. the β_{1i}) for the fitted trends in Spotfin Chub log-CPUE values over the study years. The red reference line in the right-panel of Figure 6 denotes ‘no average annual change’ across the existing years of data. Note that the right-panel confidence intervals associated with each sampling location covers the ‘no average annual change’ value (i.e. slope for Years = 0) and thus there has been no significant trend in the averaged log-CPUE values over the existing years of data for any of the sampling locations. Figure 3 displays the observed CPUE values for each sampling location along with the back-transformed linear regression fits of the log-CPUE values at each of the sampling locations (Note that since we fit log-transformed values, we need took the inverse of the natural log in order to back-transform to the original scale). It should be noted that the strongest observed negative trend was associated with the Nemo location but this trend was not a statistically significant trend ($p = 0.313$). An interesting feature of the fits to each sampling location is that the fitted 2006 log-CPUE values were all systematically smaller than the observed 2006 log-CPUE

numbers. Removing the 2006 log-CPUE values the observed trends would have been even smaller than what we are currently observing.

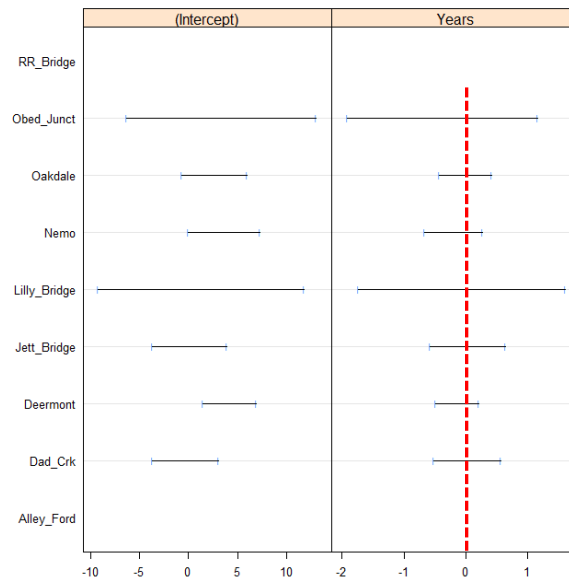


Figure 6. The 95% confidence limits on the intercepts (left-panel) and slopes (right-panel) for the fit of the linear regression model given in (6) for each of the nine sampling locations. Red reference line indicates a value of ‘no average annual change’.

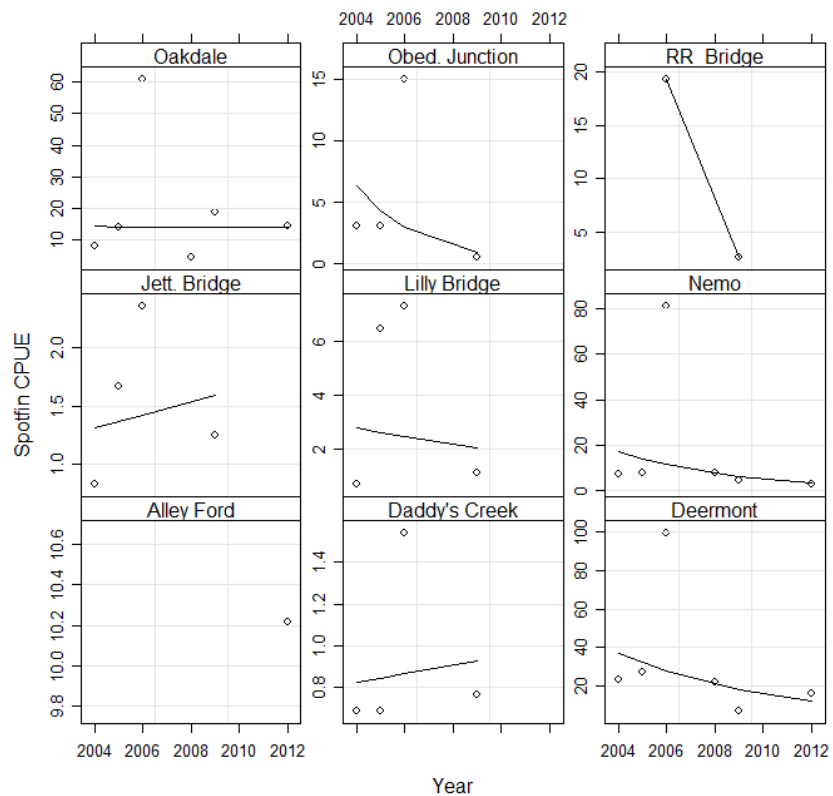


Figure 7. Spotfin Chub CPUE's with the back-transformed fits to the linear regression model in (6) for each of the nine study locations. While fits appear to suggest slight trends, when comparing the observed trend to the calculated standard errors, trends are not statistically significant. The lowest p-value is associated with Nemo $p=0.313$).

Study Area Averaged Analyses (Spotfin Chub)

This section will address the second set of questions and we will refer to these set of questions as 'study area averaged' questions since we wish to describe what is happening in general across all nine sampling locations. Before explaining the model in this section, it is important to note that for the Spotfin Chub data, since no significant trend was observed for any of the nine study locations, we will not observe a significant trend for the average trend across all nine study locations. To address the second set of questions, we can simply fit the following linear model to the pooled data (i.e. pool the data across all nine study plots):

$$\ln(\text{Count}_{ji}) = \beta_0 + \beta_1 \text{Year}_{ji} + \epsilon_{ji}. \quad (7)$$

In comparing the model in (7) to that in (6), note that there is no ' i ' subscript on the intercept or on the slope coefficient, suggesting that a common 'study-area wide' intercept and slope are fit to the entire set of data. The model in (7) does not allow for any differences in the sampling locations since we are averaging over each sampling location. The model in (7) also does not account for the fact that repeated observations on each sampling location (i.e. repeated over time) are correlated with one another. To determine if log-CPUE values differ depending upon the sampling location, we can examine boxplots of the standardized residuals from the fit of the model in (7) to see if the boxplots are systematically above and below a reference line of zero according to the sampling location. If any of the boxplots are entirely above or below zero, we need to fit a model that allows for differences among the sampling locations. Figure 8 shows the boxplots of the standardized residuals from the fit of the model in (7) according to sampling location. Note that three locations contain standardized residuals that are either all above the zero reference line or all below the reference line, thus suggesting the need to account for differences in the sampling locations. The following *linear mixed model* will be fit to the data:

$$\ln(\text{CPUE}_{ji}) = \beta_0 + \beta_1 \text{Year}_{ji} + \delta_{0i} + \epsilon_{ji}. \quad (8)$$

The model in (8) differs from the one in (7) by the random intercept component δ_{0i} . This component can be interpreted as the random deviations across sampling locations in terms of overall location abundances. So, using log-CPUE as a proxy for ‘abundance’, we observe larger a larger abundance of Spotfin Chub at the Deermont location than the other sampling locations.

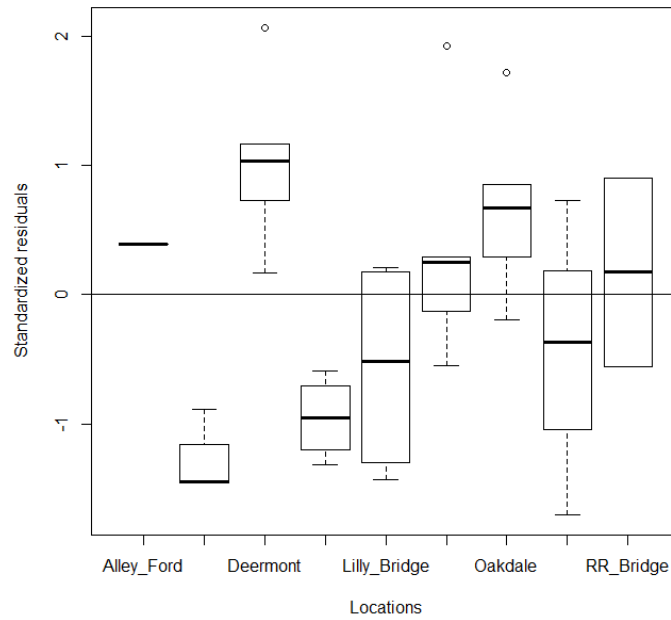


Figure 4. Boxplots of the standardized residuals from the fit of the model in (7). Note that the standardized residuals from three of the sites are either all above the zero reference line (ex. Deermont) or all below the reference line (ex. Jett_Bridge). This suggests that we need to account for differences in the sampling locations when fitting the study-area averaged model.

Fitting the model in (8) using the ‘lme’ function in R, we observe results in Table 2. The results in Table 2 are broken down into two parts: 1. Random effects and 2. Fixed effects. The random effects portion indicates the sources of variability observed in the existing data. The total variation in CPUE values is partitioned into ‘between location’ variance (i.e. $1.119^2=1.2522$) and ‘within location variability’ (i.e. $0.9489^2=0.9004$). The total variation in CPUE is then $1.2522+0.9004 = 2.1526$ and the ‘between location variability’ is estimated to account for $\left(1.2522/2.1526\right)100\%=58.17\%$ of the total variation in CPUE values. It is important to note that the model in (8) assumes equal trends across all nine sites and, as observed in Figure 7, some sites appear to have a slight positive trend over time while others have a slight negative trend

over time. Most of the variation that is currently attributed to ‘within site variation’ is primarily due to the observed 2006 CPUE values. As mentioned earlier, the predicted CPUE values for 2006 are smaller than the observed CPUE values at all sites. It is quite possible that, in actuality, the ‘between location’ variance encompasses a greater proportion than the currently observed 58.17% but until more years of data are collected and we have a better estimate of the within site variance, it is difficult to ascertain. In attempting to fit different slopes for each of the sites utilizing the linear mixed model, we were not able to obtain convergence in the estimation of the model parameters. As noted in the discussion of the Tuxedo Darter data, this phenomenon is not surprising given the limited amount of data and the very slight trends that are currently being observed.

The fixed effects portion indicates the estimated intercept and slope from the fit of the model in (8). Note that the estimated slope is -0.0892, suggesting a slight negative trend in the overall average log-CPUE values across sites. The associated p-value of 0.1909 indicates that this observed trend is not statistically significant. Thus, as expected (when considering that no significant trend was observed for any given sampling location), no significant overall trend across the study years exists for the Spotfin Chub CPUE values.

Table 2. Mixed model output from the fit of the model in (8) to the Spotfin Chub data

Random effects:

Formula: $\sim 1 \mid \text{Site}$
 (Intercept) Residual
 StdDev: 1.119 0.9489

Fixed effects: $\ln\text{CPUE} \sim \text{AbvYear}$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	2.2287	0.6267	27	3.556	0.0014
Years	-0.0892	0.0665	27	-1.342	0.1909