Statistical Consulting Services

3405 Chukar Lane, Clarkston, Washington 99403 (509) 758-8330

stat@cableone.net

• Design ***• Analyze*** ***• Report***

Status: Preliminary Report
Date: October 28, 2008
Project Title: Algae Tile Data: 2004-2007, BPA-51

I) Data Description

Multiple files containing 2004 through 2007 Tile Chlorophyll data for the Kootenai River sites designated as: KR1, KR2, KR3, KR4 (Downriver) and KR6, KR7, KR9, KR9.1, KR10, KR11, KR12, KR13, KR14 (Upriver) were received by SCS. For a complete description of the sites covered, please refer to http://ktoi.scsnetw.com.

To maintain consistency with the previous SCS algae reports, all analyses were carried out separately for the Upriver and Downriver categories, as defined in the aforementioned paragraph. The Upriver designation, however, now includes three additional sites, KR11, KR12, and the nutrient addition site, KR9.1.

There were 407, 243, 318, and 300 observations recorded for algae in 2004, 2005, 2006, and 2007, respectively. The variables contained in the tile chlorophyll data are listed in Printout 1. These were: *site (Site)*, *date (Date), replication (Rep), River Kilometers (RKM), chlorophyll a (Chla), chlorophyll a accrual rate (Chla_Acc), total chlorophyll (TotalChl),* and *total chlorophyll accrual rate (Total_Acc)*. The dates available ranged from April 25, 2004 to October 23, 2007 over the sites listed above. Replication within each site and date ranged from 3 to 10 samples. River Kilometers, added by SCS, corresponds to the distance down river for the respective site designations. All response variables are recorded (and represented here) in units of mg/m^2 for amounts and $mg/m^2/30d$ for accrual rates.

Some anomalous observations have been omitted from these analyses. These were those amount values exceeding 400 mg/m^2 and accrual rates higher than 200 $mg/m^2/30d$. These censored data points accounted for less than 1% of all the data. For a complete listing of the data set, please refer to the Algae *Auxillary Data* link on the KTOI Ecosystem WEB site at: http://ktoi.scsnetw.com .

II) Summary Statistics and Trends

Summary statistics and information on the four responses, chlorophyll a, chlorophyll a Accrual Rate, Total Chlorophyll, and Total Chlorophyll Accrual Rate are presented in Print Out 2. Computations were carried out separately for each river position (Upriver and Downriver) and year. For example, the Downriver position in 2004 showed an average *Chlorophyll a* level of 25.5 mg with a standard deviation of 21.4 and minimum and maximum values of 3.1 and 196 mg, respectively. The Upriver data in 2004 showed a lower overall average chlorophyll a level at 2.23 mg with a lower standard deviation (3.6) and minimum and maximum values of (0.13 and 28.7, respectively).

A more comprehensive summary of each variable and position is given in Print Out 3. This lists the information above as well as other summary information such as the variance, standard error, various percentiles and extreme values. Using the 2004 Downriver Chlorophyll a as an example again, the variance of this data was 459.3 and the standard error of the mean was 1.55. The median value or 50^{th} percentile was 21.3, meaning 50% of the data fell above and below this value. It should be noted that this value is somewhat different than the mean of 25.5. This is an indication that the frequency distribution of the data is not symmetrical (skewed). The skewness statistic, listed as part of the first section of each analysis, quantifies this. In a symmetric distribution, such as a Normal distribution, the skewness value would be 0. The tile chlorophyll data, however, shows larger values. Chlorophyll a, in the 2004 Downriver example, has a skewness statistic of 3.54, which is quite high. In the last section of the summary analysis, the stem and leaf plot graphically demonstrates the asymmetry, showing most of the data centered around 25 with a large value at 196. The final plot is referred to as a normal probability plot and graphically compares the data to a theoretical normal distribution. For chlorophyll a, the data (asterisks) deviate substantially from the theoretical normal distribution (diagonal reference line of pluses), indicating that the data is non-normal. Other response variables in both the Downriver and Upriver categories also indicated skewed distributions. Because the sample size and mean comparison procedures below require symmetrical, normally distributed data, each response in the data set was logarithmically transformed. The logarithmic transformation, in this case, can help mitigate skewness problems. The summary statistics for the four transformed responses (*log_ChlorA, log_TotChlor,* and *log_accrual*) are given in Print Out 4. For the 2004 Downriver Chlorophyll a data, the logarithmic transformation reduced the skewness value to -0.36 and produced a more bell-shaped symmetric frequency distribution. Similar improvements are shown for the remaining variables and river categories. Hence, all subsequent analyses given below are based on logarithmic transformations of the original responses.

Trends

Plots of the four responses across sampling date are given in Print Out 5. These plots provide both the mean trend (solid lines) and measures of variation (box plots) for each river month. For each month, vertical lines extend from the minimum value to the maximum. Blue boxes represent the data spread between the 25^{th} and 75^{th} percentiles. The trends for all four variables were similar within each year. The seasonal changes in the 2004 data seem to be minimal in June and July, while the trends in 2005, 2006 and 2007 differ and may, in some cases, increase over time.

Print Out 6 provides similar trend plots for the responses over sample sites. Lines, trends and boxes are interpreted in the same manner as the previous printout. These plots show chlorophyll production parameters to be variable over river distance with higher values occurring mainly in the Downriver sites in 2004. Starting in 2005, the nutrient addition site, KR9.1, shows higher levels of chlorophyll. This continues in 2006, but is mitigate in 2007 where chlorophyll levels are elevated at all sites.

III) Determination of Sample Sizes

When estimating the mean, the formulation for calculating sample size is:

$$n = (z*s/d)^2$$

where s, d and z are related to the variability, desired precision, and confidence levels, respectively. As was stated previously, prior transformation of the data is necessary to meet the distributional requirements for this estimation.

The desired precision was fixed at 10% of each response mean value. For the Downriver Chlorophyll a example, 10% of the overall mean is approximately 2. The confidence level, z, is obtained from a standard table of normal values, i.e.1.96 for 95% confidence. Standard deviations, s, were obtained for each response and the resulting sample size estimates computed (Print Out 7). Here, sample size estimates have be calculated for confidence levels of 90, 95 and 99%. All sample size estimates at the 95% level of confidence are within or close to the current sampling protocol of 6 replications for 2004. For the remaining years, this is also true with the exception of Chlorophyll a measurements where the sample size estimates are somewhat higher. Most notable is the Upriver section in 2005 where estimates were all larger. This may be due to more variable responses in that year. Overall, however, retaining the current level of sampling intensity (6 replications) should ensure that the desired precision of 10% of the mean value can be approximately achieved for all responses in the Upriver and Downriver categories with at least 95% confidence.

It should be noted that for all the above calculations, the resulting sample size values are preliminary. The calculations are based on limited data covering a single season. Furthermore, the actual precision of mean estimates may vary by sampling date. Therefore, these sample size values should be used cautiously for setting policy regarding future sampling protocols.

IV) Additional Analysis

An additional comparative analysis was also conducted to investigate the difference in average response (primary productivity) between the designated river positions. Print Out 8 provides a statistical comparison of the response means across the Upriver and Downriver sites. For all responses, in all years, the analyses show significant differences between the Downriver and Upriver designations. For the 2004 Chlorophyll a example, the F statistic testing for mean differences was 825.15 and the associated p-value 0.0001. Since this F statistic is very large and the p-value smaller than 0.05 (i.e. a 95% level of significance), we conclude that there was evidence of a significant difference between the average Upriver and Downriver Chlorophyll a levels.

The second section of the printout provides the model estimated means for both the Upriver and Downriver categories. Again for 2004 Chlorophyll a, the means were 2.95 and 0.17 for Downriver and Upriver, respectively. Since these means are reported on the logarithmic scale, we must untransform them to obtain usable results. This is done using the exponential or anti-log function. For the Downriver and Upriver categories this gives: exp(2.95) = 19.1 mg and exp(0.17) = 1.2 mg, respectively.

V) Year Comparisons

Print Out 9 gives mean trend lines across sample sites for the four responses in years 2004, 2005, 2006, and 2007. All responses show a general increase in magnitude over years with KR9.1 (nutrient addition site) increasing markedly in 2005 and 2006. Trends for 2007 are less pronounced, however, with higher overall measurements at most sites.

VI) Additional Remarks

1) It is important to maintain a consistent sampling protocol at each sampling date and location. This will help insure that the desired precision level is reached at all locations and times as well as facilitate the potential modeling of temporal effects later on.

2) While some data (< 1%) were potentially questionable, their omission from this report was deemed to have a negligible effect.

3) The format and structure of the 2004-2007 algae data received by SCS was variable. Maintaining a consistent data format in the future will facilitate subsequent data analyses and incorporation of the data into the relational database.