Appendix E TYC Survey Data Methods and Analysis

Section 2.4 TYC distribution

Knowledge on TYC distribution is based on the monitoring data collected from 1979 to 2000. The measured variables in the dataset include stem counts (Appendix C) and lake level in the first week of September (Figure 2.5). There are 30 years of survey data and up to 28 estimates of stem count available for a site. From these data, an occupancy level can be calculate for each of the 30 survey years (the number of occupied sites/the number of sites surveyed*100) (shown as bars in Figure 2.7) and a mean stem count calculated for each year. A single calculation of persistence during the period can be made for each site (the number of years a site is occupied/the number of years surveyed *100). It is unclear how stem counts may or may not relate to population size (see section 2.6). Furthermore, stem counts may reflect the amount of habitat available and also impacts from trampling. However, recreation use has not been quantified. As such, a multiple regression framework was not pursued because there are only two measured predictor variables.

The relationship between lake level and the absolute number of TYC sites (Figure 2.8) and the proportion of surveyed TYC sites (Figure 2.9) was examined for simple correlation. Lake level is reported to the nearest whole interfere as measured in September at the USGS Tahoe City gage (103370000). All 50 sites with survey data in Appendix C were included in the analysis. At USFS sites with enclosures, stem counts from inside the enclosures and outside have been tracked separately and these were combined for Appendix C and all analyses. No surveys were conducted in 8 (1984, 1985, 1987, 1989, 2010, or 2013). Two years (1991 and 1992) with <60% survey effort were excluded. To assess the role of ownership on occupancy, occupancy was calculated for each of three 3 ownership categories (USFS, State, private) for each survey year (Figure 2.10).

Data were analyzed using the statistical software R version 3.1.1 (R Core Team 2014). The assumption of normality required of linear regression was mildly violated as evidenced by examination of QQ plots. Therefore, a more conservative test (Spearman's rank correlation coefficient) was utilized to test the relationship between lake level and TYC occupancy. Spearman's rank is a nonparametric measure of statistical dependence between two variables (continuous or discrete) that assesses how well the relationship can be described using a monotonic function. The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. In this case, a perfect Spearman correlation of −1 would result if the relationship between number of occupied sites and lake level were perfectly linear, so a value of -0.80 indicates the relationship is very strong. Spearman's rank correlation first converts data to ranks and so reduces the likelihood of a Type 1 error. Compared to Pearson's r, another commonly used correlation coefficient, Spearman's is less influenced by outliers.

Section 2.6 Population size and persistence

Analysis in 2002 plotted mean stem counts for 29 sites (with high quality records) against their calculated persistence value (Pr) over the period from 1979 to 2000 and fit the data with a logarithmic curve (see Appendix G for Figure 13 and 14 in Pavlik *et al*. 2002). The fit was significant at $p<0.01$ using a bi-variate test and the r^2value indicated that stem counts explained 63% of the variation in persistence. Logistic regression performs a non0linear transformation on a generalized linear model but constrains the outcome to between 0 and 1, and so predicts a probability.

This analysis was repeated for the larger dataset from 1979 to 2014 for 45 sites that fit the ranking criteria. The logarithmic curve was fitted using Excel (Figure 2.11). However, analysis of frequency histograms determined that stem counts and Pr were not normally distributed and so violate the assumptions required for logistic regression analysis. Spearman's rank correlation coefficient was used to test the relationship between stem count and persistence. Spearman's converts data into ranks and compares relative order so it is less susceptible to violations of normalcy (a requirement of logistic regression). A Spearmans' $\rho=0.62$ ($p<0.01$) indicates a positive correlation between stem count and persistence.

The relationship between mean stem count of all sites for each year and lake level was analyzed using 23 years of data (Figure 2.12). No stem counts were available for 1994-1998, and 1991 and 1992 were excluded for low survey effort (<60%). A Spearman's rank correlation ($\rho= -0.69$ ($p<0.002$) on the dataset indicates a relatively strong negative correlation between stem count and lake level.

Section 2.7 Ranking TYC sites for conservation

The following formula was used to calculate each site viability index (SVI = Ra + (-1*CoVar) + Pr. To be ranked, sites had to have 10 years of consecutive survey data and have at least three recorded stem counts. Of the 55 survey sites, 45 sites met the criteria. This analysis used a slightly different calculation than the 2002 analysis (CS2002). Pr for a site was weighted by the proportion of years that the site was surveyed to reduce the importance in sites that have only been surveyed in the most recent years. For instance, Sugar Pine Point was occupied in all of the last 12 years and so had an un-weighted Pr of 100, but was only surveyed for 40% of the survey years, so the weighted SVI became -47 instead of 13 and, thus, the rank became Ephemeral. Weighting the Pr also led to more definitive breaks for defining ranks and lead to calculated SVI values of 103 to 2 for ranked sites that is very similar to an intuitive 100 point scale. Calculated SVIs ranged from a high of 103 to -83.

In order to determine meaningful breaks among site rank classes and provide a repeatable methodology, a form of cluster analysis called a Jenk's natural breaks optimization was conducted for all sink rank values using free software "Real Statistics for Excel" software & methodology. This method partitions the range of site rank values into contiguous classes to maximize the variation between classes and minimize the variation within the group. To verify how many site rank classes were warranted, optimization was conducted for seven partitions (2 to 8) and a goodness of variance fit (GVF) was calculated for each partitioning

for all iterations of site rank values, then the relative percentage between partitions was calculated to determine the best fit relative to the next smallest partition (Figure 1).

The absolute change in the GVF between each partition was plotted and shows that it approaches one as the number of breaks increases (Figure 1). The percent change between each successive partition was plotted on a secondary axis (Figure1). Relative change was calculated as the absolute change in GVF between partitions divided by the absolute change in GVF between the next largest partitions. Relative change in GVF shows that the most substantial breaks—those that provide the best fit relative to the next largest break—are at at 3 (65% increase in GVF relative to next break) & 5 (90%). Therefore, five breaks were selected that corresponded to TYC site ranks of Core, High, Medium, Low, and Ephemeral (Table 1). Section 2.7 discusses how these breaks are biologically meaningful for TYC conservation.

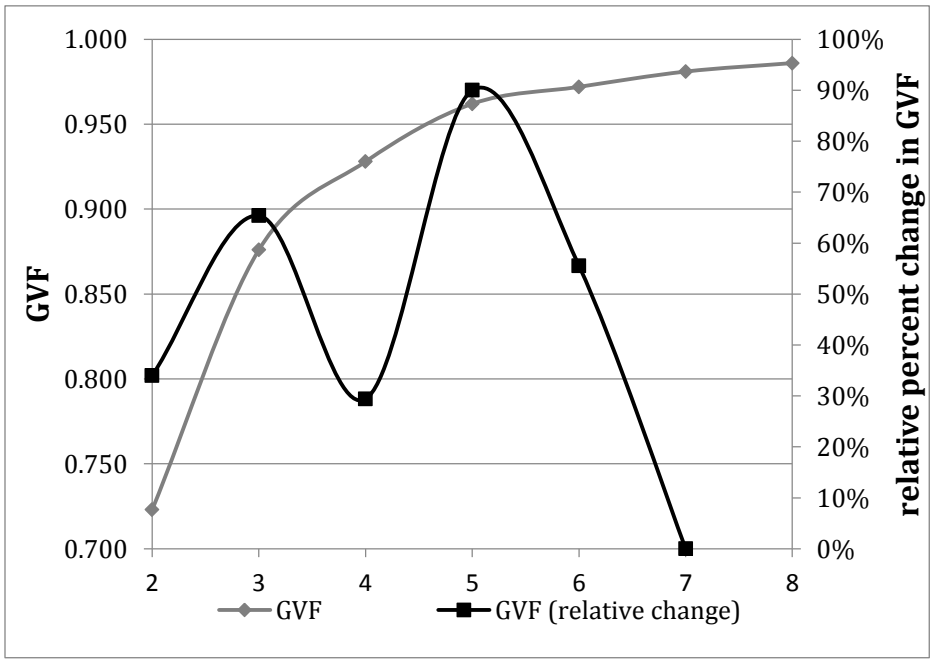

Figure 1. Jenk's natural breaks optimization of goodness of variance fit (GVF) and the relative percent change in GVF between partitions of 2 to 8 breaks

Table 1. SVI values and rank classes for best fit partitioning of SVI values (5 breaks)

break	*lower SVI*	*upper SVI*	*count*	*rank class*
1	-83	-52	6	Ephemeral
2	-47	14	11	Low
3	-6	19	11	Medium
4	34	60	11	High
5	70	103	6	Core
GVF = 0.962, relative change from next largest partition = 90%				

A Kruskal-Wallis test was used to evaluate differences in mean SVI, persistence (Pr), relative abundance (Ra), Coefficient of variation (CoVar), and stem counts of TYC survey sites for each of the five ranking categories (Table 2.2). The Kruskal-Wallis is a non-parametric analogue to a one-way ANOVA and is used when non-normally distributed data from repeated sampling would violate the assumptions of ANOVA. The test was significant for all variables at $p<0.01$. Post-hoc analysis ($\alpha = 0.5$) revealed significant differences among pairs. Within a column sites were assigned different letters.

The Wilcoxon signed rank test was used to test if ownership played a role in how sites ranked. It is a non-parametric analogue to the paired t test which determines if the ranked population means of the two compared groups differ. When there are more than two groups it is equivalent to the Kruskall Wallis. The mean calculated site viability index (SVI), persistence (Pr), and mean stem count of TYC survey sites from 1979 to 2014 with public/mixed ownership were not significantly different from privately owned sites ($\alpha = 0.5$).

The mean persistence (Pr) and stem count of 45 TYC sites surveyed from 1979 to 2014 associated with a creek, other outflows, or no water flow were evaluated with a Kruskall Wallis test because analysis of frequency histograms determined that stem counts and Pr were not normally distributed and so violate the assumptions required for ANOVA. Within a column, sites with different letters are significantly different at $\alpha=0.05$.

References

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *http://www.R-project.org/*.

Real Statistics for Excel software. Accessed September, 2015. *http://www.real-statistics.com/multivariate-statistics/cluster-analysis/jenks-natural-breaks/*.