Draft Report

# Rivers and Lakes Bring the Boys to the Yard: Assessing the Effect of Riparian Zones on Taxonomic Richness

## Introduction

Throughout the mythologies of ancient civilizations, rivers are often portrayed as sacred lifelines and mystical forces that breathe life into the land. From the Nile of Egypt to the Ganges in India, rivers shaped the very foundation of human development. The route a river carves through a landscape not only guides the path of its water, but also dictates the types of vegetation that grow alongside it, the diversity of animals that depend on it, and the feasibility of long-term human settlement. The majority of the world’s great cities were born beside rivers, not merely for commerce, which could be achieved from any coast, but for one elemental advantage: rivers carry fresh water. Drinkable water. Or at least, they used to. In today’s world, the demand for freshwater is greater than ever. Global population growth, expanding agricultural demands, urban sprawl, and climate change are placing unprecedented stress on aquatic ecosystems. Historically, human progress has been closely tied to our ability to manipulate the quantity of water available through irrigation systems, dams, aqueducts, and other technologies. But as environmental pressures grow, we are now forced to reckon with a second, equally critical dimension: water quality. Just as our ancestors learned to interpret the presence of certain plants and animals as signs of viable water sources, modern ecologists use biological indicators to evaluate the health of freshwater systems. Among the most widely used of these indicators are benthic macroinvertebrates; organisms that live on the bottom of rivers, lakes, and streams and include insects, crustaceans, mollusks, and worms. These organisms are uniquely valuable for biomonitoring due to their varied tolerance to pollutants, their limited mobility (which ties them to local conditions), and their life cycles, which span seasons or years. One particularly important group of macroinvertebrates used to indicate good water quality is the EPT taxa; Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies). These taxa are generally sensitive to environmental disturbances, and their presence in high abundance is often correlated with clean, well-oxygenated water. In contrast, a dominance of more pollution-tolerant organisms such as oligochaete worms, leeches, and midge larvae (Chironomidae) is typically a sign of poor water quality and habitat degradation (McDonald, Mullins, and Lewis, 1991). However, water quality is not influenced solely by what occurs within the streambed. The surrounding environment, particularly riparian zones, plays a critical role. Riparian zones are the transitional areas between terrestrial and aquatic ecosystems, typically encompassing the vegetation and soils adjacent to rivers and streams. Though narrow in spatial extent, riparian zones provide disproportionate ecological benefits. They act as natural buffers, filtering out sediments, nutrients, and pollutants before they reach the water body. They also stabilize stream banks, reduce erosion, moderate water temperature through shading, and provide crucial habitat for a wide range of terrestrial and aquatic organisms. As such, the integrity of riparian zones is often used as a secondary, yet equally important, indicator of watershed health (Sabo et al., 2005). Unfortunately, riparian zones are frequently the first casualties of human development. Agricultural expansion, urbanization, logging, and road construction often involve the removal or fragmentation of riparian vegetation. When riparian zones are degraded, their ability to support both terrestrial and aquatic biodiversity diminishes, and the systems they protect become more vulnerable to pollution, sedimentation, and thermal stress. Consequently, monitoring the type and abundance of riparian vegetation can yield valuable insights into the overall ecological condition of a watershed, and by extension, its capacity to support a diverse and healthy macroinvertebrate community. With these connections in mind, our project seeks to investigate the relationship between riparian zone condition and freshwater invertebrate diversity. Specifically, we aim to evaluate whether the composition and abundance of riparian plant cover is associated with the composition and abundance of benthic macroinvertebrates, with an emphasis on EPT taxa as indicators of ecosystem health. By exploring these patterns, we hope to contribute to a better understanding of how human land use and vegetation management practices may impact water quality through both direct and indirect ecological pathways. To conduct this analysis, we will use data from the Klamath Network Lakes Monitoring Data Package (2013–2019), which includes biological and ecological monitoring data collected across three national parks: Crater Lake, Lassen Volcanic, and Redwood National Parks. From this dataset, we will focus on three specific tables: InvertSpecies.csv, which documents the abundance of macroinvertebrates at various sampling plots, RiparianCover.csv, which describes the vegetation characteristics of corresponding riparian zones, and WaterQuality.csv which includes data on physical and chemical water parameters at each sampling plot. Integrating this dataset will allow us to account for potential covariates or confounding variables, such as dissolved oxygen, pH, or turbidity, that may influence macroinvertebrate populations independently of vegetation conditions. Using this data, we will assign a standardized “vegetation score” to each plot based on the richness and type of plant cover recorded. This score will serve as a proxy for riparian zone quality. We will then analyze the relationship between this vegetation score and macroinvertebrate diversity, particularly the presence of sensitive EPT taxa, using statistical methods in RStudio. To determine the strength and direction of relationships between variables, we will conduct correlation tests (Pearson or Spearman, depending on normality assumptions). To model the predictive influence of vegetation on macroinvertebrate abundance, we will use linear regression analysis. These statistical approaches will be complemented by clear and informative data visualizations created with ggplot2, including scatterplots, box plots, and bar graphs to communicate trends and variation across sites. This holistic approach will allow us to better interpret the ecological relationships we uncover and provide more robust insights into freshwater health within these protected park systems.

## Data Exploration

All data for this project comes from the Klamath Lakes Monitoring Data Package 2013–2019, originally compiled by the U.S. Department of the Interior and now publicly available through Data.com. This dataset includes ecological monitoring records from Crater Lake National Park (NP), Lassen Volcanic NP, and Redwood National and State Parks (NSP). However, for the purposes of our study, we will exclusively use data collected at Crater Lake National Park, narrowing the geographic focus to improve consistency and data quality. The three files central to our analysis are InvertSpecies.csv, RiparianCover.csv, and WaterQuality.csv. These files contain complementary data and share key identifying columns, including Lake\_Code, Start\_Date, and Park\_Name—which will allow us to merge the datasets and compare biological, physical, and vegetative conditions at each sampled water body. The InvertSpecies file contains records of benthic macroinvertebrate communities sampled in lakes and ponds. It includes fields such as Taxon, Order, Count, and Life\_Stage. For this study, we will focus on the orders Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies), which are collectively known as EPT taxa. These macroinvertebrates are widely recognized as indicators of healthy freshwater ecosystems, as they are generally sensitive to pollution and habitat degradation. The RiparianCover file provides plot-level data on vegetation type and structure in riparian zones, such as Big\_trees\_ArealCover, Woody\_shrubs\_saplings\_GroundCover, Canopy\_Type, and Barren\_AreaCcover. These variables will help us estimate the ecological integrity and potential human impact on surrounding shorelines. Each plot corresponds to a specific lake, allowing us to associate vegetation conditions with macroinvertebrate presence in those water bodies. The WaterQuality file offers insight into the abiotic conditions of each lake, including key measurements like Temperature\_C, pH, Turbidity, Dissolved\_oxygen, and Specific\_conductivity. These parameters may contribute to or constrain macroinvertebrate diversity and could help contextualize any patterns observed between vegetation and aquatic biodiversity. While all three files have a relatively large number of rows (RiparianCover.csv: 1,680; InvertSpecies.csv: 3,299), some fields contain missing values, particularly with respect to sampling dates and environmental measurements. To account for this, we will filter for data collected in 2018, a year with strong representation across datasets, and group observations by plot. This will allow us to match vegetation conditions with corresponding macroinvertebrate counts and water quality measurements. From RiparianCover.csv, we are particularly interested in a subset of vegetation metrics that reflect both habitat structure and disturbance: Big\_trees\_ArealCover, Woody\_shrubs\_saplings\_GroundCover, and Barren\_ArealCover. These will form the basis of a vegetation condition score used later in the analysis. Meanwhile, from InvertSpecies.csv, we will extract insect counts and filter specifically for EPT taxa, which we define as “significantly present” when observed in quantities of at least 10 individuals per plot. Lastly, WaterQuality.csv will be used to examine physical and chemical characteristics of each lake. These variables will not only support broader ecological interpretations but also offer an additional layer of explanatory context for understanding macroinvertebrate patterns. By combining these three datasets with consistent identifiers and a focused temporal and spatial scope (Crater Lake, 2018), we can begin exploring the relationships among riparian cover, water quality, and macroinvertebrate biodiversity with clarity and coherence.

### Data Visualizations

Data Preparation

# Load libraries  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

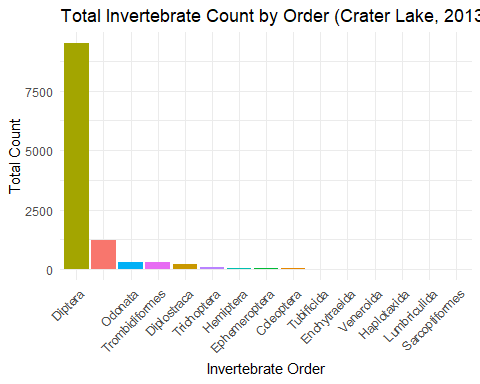
The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)  
  
# Load datasets  
invert\_species <- read.csv("InvertSpecies.csv")  
riparian\_cover <- read.csv("RiparianCover.csv")  
water\_quality <- read.csv("WaterQuality.csv")  
  
# Filter for Crater Lake and 2018  
invert\_species\_filtered <- invert\_species %>%  
 filter(Park\_Name == "Crater Lake National Park" & Start\_Date >= "2018-01-01" & Start\_Date <= "2018-12-31")  
  
riparian\_cover\_filtered <- riparian\_cover %>%  
 filter(Park\_Name == "Crater Lake National Park" & Start\_Date >= "2018-01-01" & Start\_Date <= "2018-12-31")  
  
water\_quality\_filtered <- water\_quality %>%  
 filter(Park\_Name == "Crater Lake National Park" & Start\_Date >= "2018-01-01" & Start\_Date <= "2018-12-31")  
  
# Merge datasets  
merged\_data <- invert\_species\_filtered %>%  
 inner\_join(riparian\_cover\_filtered, by = c("Lake\_Code", "Park\_Name", "Start\_Date")) %>%  
 inner\_join(water\_quality\_filtered, by = c("Lake\_Code", "Park\_Name", "Start\_Date"))

Source: [Article Notebook](https://mckenna-cooper.github.io/lab_7_final/index.qmd.html)

\*\*Note: Although it is stated above and in Preliminary Methods that we would only be evaluating Carter Lake National Park in the year 2018, this was not feasible for the drafted visualizations. We had to pivot to a different proposal idea due to the elimination of data this past weekend, and need more time to narrow down how to visualize with these specific data sets. 1.

invert\_species %>%  
 filter(  
 Park\_Name == "Crater Lake National Park",  
 Start\_Date >= "2013-01-01", Start\_Date <= "2019-12-31",  
 !is.na(Order),  
 !is.na(Count)  
 ) %>%  
 group\_by(Order) %>%  
 summarize(Total\_Count = sum(Count, na.rm = TRUE)) %>%  
 ggplot(aes(x = reorder(Order, -Total\_Count), y = Total\_Count, fill = Order)) +  
 geom\_col(show.legend = FALSE) +  
 labs(  
 title = "Total Invertebrate Count by Order (Crater Lake, 2013-2019)",  
 x = "Invertebrate Order",  
 y = "Total Count"  
 ) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

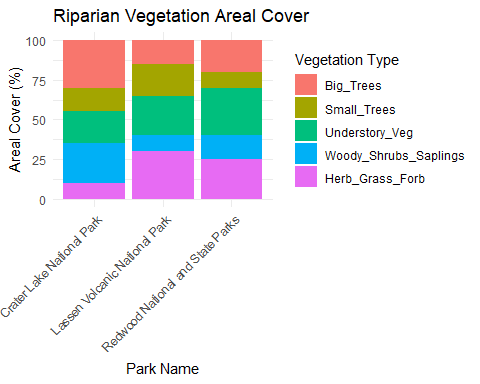


Source: [Article Notebook](https://mckenna-cooper.github.io/lab_7_final/index.qmd.html)

# Load necessary library  
library(ggplot2)  
  
# Create a sample dataset  
data <- data.frame(  
 Park\_Name = c("Redwood National and State Parks", "Lassen Volcanic National Park", "Crater Lake National Park"),  
 Big\_Trees = c(20, 15, 30),  
 Small\_Trees = c(10, 20, 15),  
 Understory\_Veg = c(30, 25, 20),  
 Woody\_Shrubs\_Saplings = c(15, 10, 25),  
 Herb\_Grass\_Forb = c(25, 30, 10)  
)  
  
# Melt the data for ggplot  
library(reshape2)

Warning: package 'reshape2' was built under R version 4.4.3

data\_melted <- melt(data, id.vars = "Park\_Name",   
 variable.name = "Vegetation\_Type",   
 value.name = "Areal\_Cover")  
  
# Plot stacked bar chart with tilted axis labels  
ggplot(data\_melted, aes(x = Park\_Name, y = Areal\_Cover, fill = Vegetation\_Type)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(title = "Riparian Vegetation Areal Cover",   
 x = "Park Name",   
 y = "Areal Cover (%)",   
 fill = "Vegetation Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://mckenna-cooper.github.io/lab_7_final/index.qmd.html)

## Preliminary Methods

To begin our analysis, we will first join the three datasets—InvertSpecies.csv, RiparianCover.csv, and WaterQuality.csv—using shared identifying columns: Lake\_Code, Park\_Name, and Start\_Date. We will then filter the merged dataset to include only Crater Lake National Park and data collected in 2018, to maintain consistency in time and location. Next, we will divide our dataset into two groups of lakes based on the presence or absence of sensitive macroinvertebrate orders: Ephemeroptera, Plecoptera, and Trichoptera (collectively referred to as EPT taxa). These taxa serve as bioindicators of high water quality and intact riparian zones. A lake will be categorized as “EPT-present” if at least 10 individuals from any of the three orders are found. Lakes without EPT taxa or with fewer than 10 individuals will be categorized as “EPT-absent.” Once we have established these two groups, we will assess potential relationships between riparian vegetation characteristics and the presence of EPT taxa. Key vegetation variables from RiparianCover.csv will include: Big\_trees\_ArealCover

Woody\_shrubs\_saplings\_GroundCover

Barren\_ArealCover

We will explore creating a vegetation integrity index by standardizing and summing Big\_trees\_ArealCover, Woody\_shrubs\_saplings\_GroundCover, and inverse Barren\_ArealCover scores. We will also consider select abiotic water quality variables from WaterQuality.csv—such as Temperature\_C, pH, and Dissolved\_oxygen—to help contextualize our findings or identify confounding influences. For statistical analysis, we will choose appropriate tests based on the type of variables being compared:

* For categorical comparisons (e.g., EPT presence vs. vegetation category), we may use a Chi-squared test.
* For numerical comparisons (e.g., EPT presence vs. percent vegetation cover), we will consider a T-test or ANOVA.
* If relationships appear linear, we may explore correlation coefficients or linear regression models to assess strength and direction.

To support our findings, we will use ggplot2 in R to create clear and informative visualizations—such as boxplots, bar charts, or scatterplots—highlighting any significant trends or differences between the EPT-present and EPT-absent lake groups. Finally, missing data and uneven sampling will be carefully handled—either through exclusion of incomplete rows or imputation strategies, depending on the nature and extent of the gaps.