Development of a Watershed Health Index for Washington State

A Thesis
Presented in Partial Fulfillment of the Requirements of the
Degree of Master of Science
With a
Major in Environmental Science
in the
College of Graduate Studies
University of Idaho
By
Yu-Li Lin (Peter)

Approved by:
Major Professor: Zachary Kayler, Ph.D.
Committee Members: Xiaogang Ma, Ph.D.; Scott Collyard, M.S.
Department Administrator: Lee Vierling, Ph.D.

May 2022

Abstract

The health of freshwater systems has been in decline and human activities have been known to be the cause of the degradation of freshwater quality. To maintain our freshwater resources, watershed health assessment efforts are needed. To retain consistency and comparability between watersheds health assessments, construction of a Watershed Health Index (WHI) is an effective method to assess and compare watershed health. Several state agencies around the U.S. have developed WHIs to assess watershed health within their state. In this study, we aimed to develop a WHI for Washington State with past watershed data through data-based approaches. We used the metrics B-IBI as response variables and the watershed health indicators as the predictor variables in Partial Least Square (PLS) regression and Boosted Regression Tree (BRT) to generate relative influences of each indicator on the variability of B-IBI, and compiled the relative influences into weight factors to calculate the final WHI. The results showed low R^2 (between 0.05 to 0.30) for the PLS regressions, and the two methods (BRT and PLS) resulted in different weight factors in the WHI. We suspect that the data structure and sample size may have contributed to the low R^2, since most samples were non-continuous measurements of watershed indicators, or "snapshots", which may capture extreme events that may not represent the actual health condition of the watersheds.

Acknowledgement

I would like to express my greatest appreciation for my major professor and the members of my committee. My major professor Dr. Zachary Kayler has showed endless support to help me finish this project, I could not have completed this study without his knowledge and inspiration. Dr. Xiaogang Ma from the College of Engineering Computer Science Department was extremely patient with me and always provided great guidance to help this project move forward, this project would not have been completed without his valued input and advise. Scott Collyard from Washington State Department of Ecology provided extremely valuable knowledge and information on the dataset used in this study, he also managed to find time for my project meetings in his busy work schedule, which I am very grateful of.

I would also like to thank the faculty and staff of the Environmental Science Department, Soil and Water Systems, and College of Graduate Studies for their patience and help throughout my time at the University of Idaho. I also thank the Washington State Department of Ecology for providing the dataset and useful resources used in this study.

This project was supported by funds from Kayler Lab and Soil and Water Systems, I couldn’t have completed this project without the funding support.

Dedication

I would like to thank my wife and my family for their endless support, I could not have done it without them. My wife Hannah took care of little things in life to help me focus on the project and always treated me with patience and positive energy. My parents, Ben and Sunny, helped me emotionally and financially throughout my time in the U.S. and always supported what I wanted to do in life. I would also like to thank my good friend Bob, who passed away last year, for his friendship and support. I wouldn’t have been able to get used to life here in the U.S. without Bob and my friends at Drift Office.

Table of Contents

Abstract ……………………………………………………………………………...ii

Acknowledgements…………………………………………………………………iii

Dedication……………………………………………………………………………iv

Table of Contents……………………………………………………………………v

List of Tables…………………………………………………………………………vii

List of Figures………………………………………………………………………viii

Chapter 1: Introduction………………………………………………………………1

1.1 Problem Statement………………………………………………………1

1.2 Purpose of Study…………………………………………………………4

1.3 Understanding Watershed Health Assessments…………………………5

Chapter 2: Review of Watershed Health Indices………………………………………7

2.1 WHI review………………………………………………………………7

2.2 Index structure and contents of WHI……………………………………17

Chapter 3: Development of Watershed Health Index for WA…………………………21

3.1 Introduction………………………………………………………………21

3.2 Material and Methods……………………………………………………24

3.3 Results……………………………………………………………………35

3.4 Discussion……………………………………………………………….…50

Chapter 4: Conclusion……………………………………………………………....57

References……………………………………………………………….……………61

Appendix………………………………………………………………………………67

Appendix A: PLS regression for samples with B-IBI >75……………………67

Appendix B: PLS regression for samples with B-IBI < 25……………….….68

Appendix C: Python Code for Data Extraction……………………………………69

Appendix D: Python code for data normalization and data imputation………...73

Appendix E: R code for PLS analysis…………………………………………74

Appendix F: R code for BRT analysis…………………………………………83

List of Table

Table 1: Summary of the sub-indices in the WHI…………………….…………………18

Table 2: PCA Loadings of each stream health indicator…………………………………36

Table 3: Percent influence of WHI sub-indices on final WHI score……………………..46

Table 4: Percent influence of individual indicators on WHI sub-indices…………...……46

Table 5: Calculated percent influence of individual indicators on final WHI……..……..47

Table 6: Percent influence of WHI sub-indices on final WHI score..…………………....47

Table 7: Percent influence of individual indicators on WHI sub-indices……………..……48

Table 8: Calculated percent influence of individual indicators on final WHI………..…..48

Table 9: The mean, standard deviation and the coefficient of variation of indicators..….49

Table 10: Comparison of percent influence of watershed health indicators…………..…51

Table 11: Result of WHI score vs B-IBI score……………………………………….....52

Table 12: The relative influence of B-IBI metrics with isolated data………………..…..55

List of Figures

Figure 1: The conceptual figure of WHI components..4

Figure 2: The conceptual figure of the WHI structure proposed by U.S. EPA………………..8

Figure 3: The conceptual figure of the main WHI structure in WHAF……………..……10

Figure 4: The conceptual figure of the main WHI structure in WSHI………………...…14

Figure 5: The conceptual figure of the main WHI structure in IWAF………………...…16

Figure 6: Machine learning logic tree……………………………………………...…28

Figure 7: Conceptual figure of the general structure of the WHI………………………..31

Figure 8: Conceptual pipeline of creating and using our WHI…………………………….32

Figure 9: Illustration of the linear scoring scale of the initial conversion...……………......35

Figure 10: PCA biplot with loadings………………………………………………..........37

Figure 11: Result of BRT expressed the relative influence..39

Figure 12: Result of BRT expressing the relative influence of stream health indicators...41

Figure 13: Result of PLS expressed in relative influence………………………………..…...42

Figure 14: Result of PLS expressing the relative influence of indicators on B-IBI….......45

Figure 15: Radar graph of WHI structures with original data…………………………....53

Chapter 1: Introduction

1.1. Problem Statement

Freshwater resources are crucial to humans and terrestrial life; we depend heavily on the ecosystem services provided by freshwater systems (Revenga, 2000). The health of freshwater systems has been in decline, and human activities have been known to cause degradation of freshwater quality, with stream systems within watersheds being the most vulnerable to urban development and agricultural activities (Allan, 2004; Hascic and Wu, 2006). Freshwater resources around the world are experiencing accelerated decline in quality and quantity, raising the concerns of freshwater availability (Wetzel, 1992). Previous studies using satellite imagery identified shifts in global water storage due to human impact and global warming, predicting potential freshwater resource insecurity (Rodell, 2018). By 2050, a population shift of 3 billion into urban areas is predicted, which would drastically increase water demand and put current urban water supply under immense pressure (Robert, 2011). In addition to water demand, contamination of freshwater sources and destruction of watersheds can also lead to the lack of freshwater availability. Contaminated water sources tend to require a much longer time period for cleaning and restoration, suggesting preservation would be far more beneficial and economical than remediation (Peters and Meybeck, 2000).

Watersheds impacted by human activities tend to experience symptoms such as altered water chemistry, physical habitat degradation, and reshaped biotic communities (Walsh et al, 2005). Alteration of habitats resulted by human activities, including modification of landscapes and vegetations in watersheds, not only alters the water quality, but also alters the hydrology of the systems (Peters and Meybeck, 2000). For instance, increased impervious surface cover tends to generate higher runoffs and reduce water storage during precipitation events (Shuster, 2005), thus increasing the risk of flooding and other hydrological concerns.

Governments around the world realize the importance of freshwater resource preservation and have implemented programs to reduce human impact. The U.S. Environmental Protection Agency (U.S. EPA) required states to adopt water quality standards that were used to assess water systems depending on the designated use as required by the Clean Water Act of 1972 (U.S. EPA, 2013). Sites that performed poorly would receive special attention from the state agencies and potentially require restoration to meet the water quality standards. EPA also suggested states to implement a form of watershed health assessment framework to monitor the health of watersheds (U.S. EPA, 2012). Preserving and maintaining the health of current watershed and freshwater systems is becoming critical to the well-being of current and future human society. Monitoring and management of these systems plays an important role in maintaining water quality and the integrity of our water resources.

To better manage freshwater resources, it is crucial to understand and identify indicators of environmental health. Environmental health indicators are measurements that signal the health of ecosystems and are often used for assessing health of environmental systems without the need of measuring all aspects of complex systems (Rice, 2003). A useful environmental health indicator needs to be measurable, quantifiable, understandable and linkable to environmental system health (IMR, 2004). Management of freshwater system health requires understanding indicators of environmental health and using the indicators for watershed health assessment. For instance, benthic community diversity is a common candidate as a stream health indicator, due to the benthic community’s sensitivity to environmental change within freshwater systems (Wetzel, 2001). When encountering environmental stress, benthic macroinvertebrate communities tend to experience diversity loss and tolerant species domination (Dudgeon et al., 2006). Benthic macroinvertebrates serve as a major food source for other vertebrates in streams, thus heavily influencing the health of other organisms in the system as well (Rosenberg and Resh, 1993). Implementing important indicators into an assessment framework to assess watershed health is crucial, as a good assessment framework can allow us to better understand watersheds and allows us to act as needed to protect the integrity of our freshwater systems.

Watershed health indicators can be categorized in many ways. Washington State used three major categories for indicators that describe watershed health: physical habitat, water chemistry, and biological indicators (Larson et al., 2019). As presented in Figure 1, individual watershed health indicators are divided into the three major categories and each indicator can influence the overall health of the watershed.

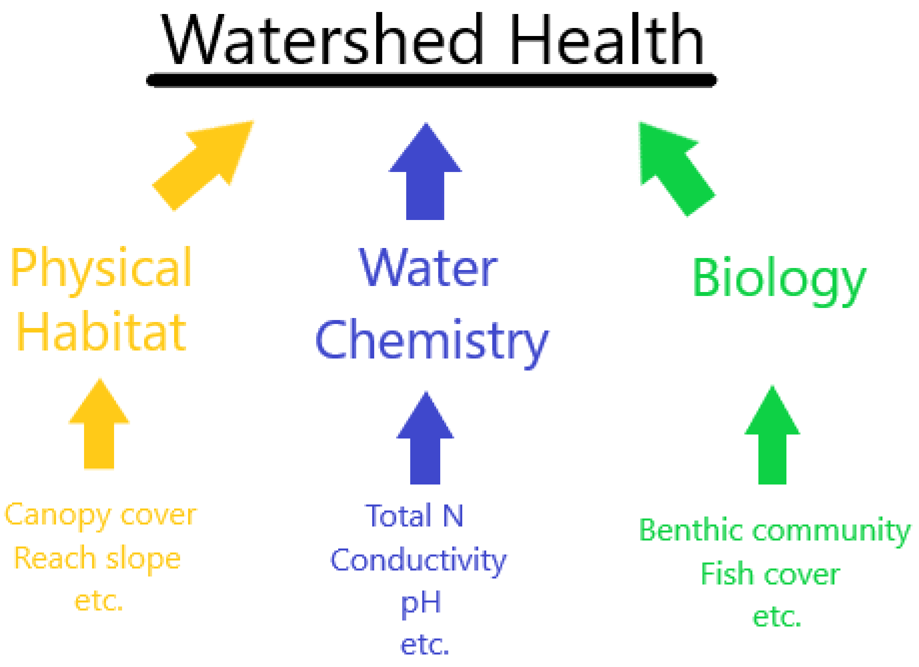


Figure 1. The conceptual figure of the components that contribute to watershed health. Individual watershed health indicators within the three categories ultimately influence the overall health of the watershed.

1.2. Purpose of Study

The purpose of this study is to explore the existing frameworks for assessing watershed health and introduce a more data-driven watershed health assessment tool for the State of Washington. The goal of my research is to develop a watershed health index that minimizes human bias and heavily relies on data analysis to construct the structure of the index. We will examine the methods and procedure of watershed health assessment frameworks constructed by the Environmental Protection Agency (EPA) and several state agencies that implemented watershed health assessments. We will attempt to use different data analytic methods to construct the watershed health index for comparison, and to understand what is needed to improve the accuracy of our watershed health index for future iterations.

1.3. Understanding Watershed Health Assessments

Watershed based monitoring and assessment is known to be a good method for understanding freshwater quality and other environmental health factors, as watershed health reflects the overall health of complex systems within the watershed (Momenian et al., 2018). The health of watershed is known to represent ecological system functions within the watershed, which is a good measurement of ecosystem function and environmental conditions. By measuring watershed health, we would also get an insight on freshwater quality and the general health of the system. Therefore, developing a watershed health assessment tool can help with the management of freshwater systems by providing watershed-scale health assessment that can be used to make management decisions within the watershed. Watershed health assessment has gained popularity among ecologists and management agencies due to the inclusion of various ecological attributes in the integrated assessment (Hoque et al., 2012). However, developing a universally applicable watershed health assessment is extremely difficult due to the complex environmental factors that are region-specific, which may require individual assessment metrics for every geographical area (Tirkey et al., 2013). Due to this reason, individual states would need to have a watershed health assessment tool for their own geographical region.

However, budget and resource constraints are challenges for local agencies to collect data for broader monitoring of watershed health (Larson et al., 2019). Furthermore, there are no standard criteria for evaluating watershed health to allow for comparison of health between different watersheds (Fleming, 1999). The purpose of a watershed health assessment would be to collect only the data that reflects watershed health, which would allow for a more efficient use of agency resources and easier assessments. Producing health assessment results that are comparable between watersheds can also simplify and reduce resource needs for agencies when trying to do watershed health assessments. To address the issue of watershed health comparability, a watershed health assessment would need to be capable of generating results that allows for comparison between watersheds within a state.

Chapter 2: Review of Watershed Health Indices

2.1.1. Overview of Watershed Health Assessment

The assessment of watershed health requires a comparable, and consistent metric that summarizes the health condition of the watershed. An inclusive index system would be a great approach to retain consistency and comparability of watershed health assessments. An inclusive index for watershed health assessment was commonly known as a Watershed Health Index (WHI). Construction of a WHI is an effective method to assess and compare watershed health (Ahn and Kim, 2017). A WHI is an assessment score aggregated from a compilation of measurable, comparable, and consistent ecological information that summarizes the primary attributes of a watershed's health condition (U.S. EPA, 2012).

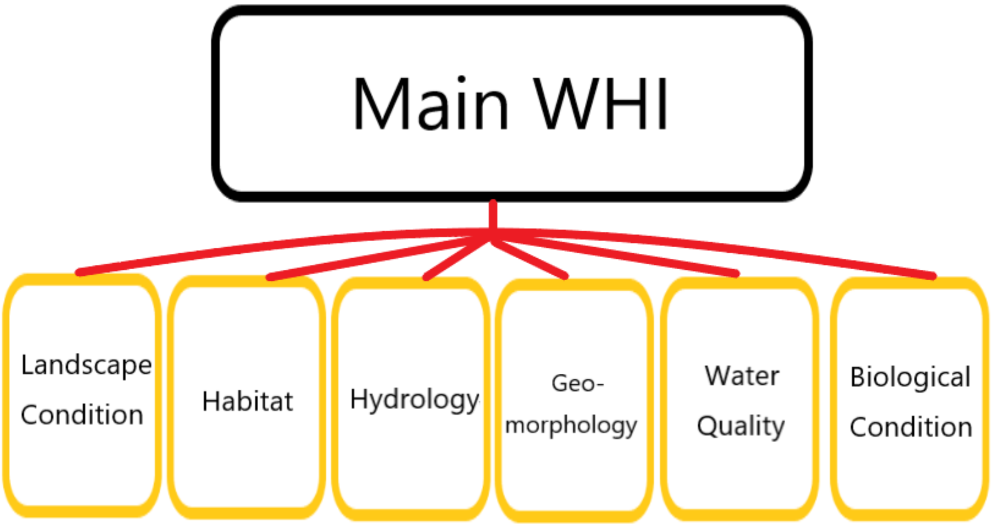


Figure 2. The conceptual figure of the WHI structure proposed by U.S. EPA, the main WHI consists of 6 sub-indices with health scores that are aggregated to produce a health score for the main WHI.

The United States Environmental Protection Agency (U.S. EPA) provided guidelines on the construction of a WHI, which included six essential ecological attributes for watershed health assessment: landscape condition, habitat, hydrology, geomorphology, water quality, and biological condition (U.S. EPA, 2012). The EPA suggested that ecological indicators in each attribute should be used to develop a sub-index per attribute, and the six sub-indices would be combined to produce a comprehensive watershed health index score. The index values were relative and not absolute, meaning that the index proposed by the U.S. EPA was only to be used for comparisons between watersheds but not treated as an absolute health score (U.S. EPA, 2012).

The guidelines provided by the U.S. EPA may not all be applicable to every region and state. Therefore, the development of a WHI should consider the area to be assessed and use data within the area for analysis. To better understand the methods used in developing a WHI, past efforts on WHI development by state agencies needs to be reviewed. By understanding past approaches we will learn insights on the advantages and disadvantages of past methods. The main questions to address are: (1) what are the critical ecological/environmental indicators to be included in the analysis, (2) what it the general structure of a WHI, (3) what are the best data exploration approaches, and (4) what are the capabilities and limitations of a WHI.

2.1.2 WHI review: Minnesota State

The Minnesota Department of Natural Resource implemented the Watershed Health Assessment Framework (WHAF) to assess health of watersheds within the State of Minnesota. In WHAF, there were five sub-indices that would each assess and produce a sub-index score for an ecological attribute of the watershed, and the 5 attributes implemented as sub-indices were taken from the ecological attributes suggested by the U.S. EPA. The five attributes include biology, connectivity, geomorphology, hydrology, and water quality. The sub-index scores is be compiled to generate a watershed health score for the watershed, scaled from 0 to 100 with 100 being the healthiest (Minnesota DNR). The main approach for aggregating sub-indices scores into the main index score was to produce a mean score by taking the average of the sub-index scores.

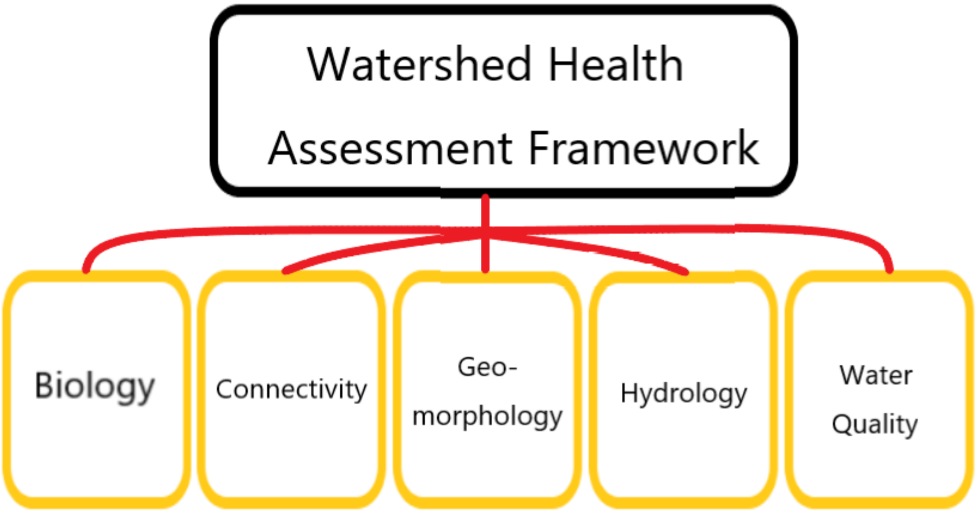


Figure 3. The conceptual figure of the main WHI structure in WHAF, the main WHI consists of 5 sub-indices with health scores that are aggregated to produce a health score for the main WHI.

The biological sub-index was broken into four components with each generating component scores: terrestrial habitat quality, stream species quality, animal species richness, and at-risk animal species richness. The terrestrial habitat quality component was generated based on five ecosystem-specific habitat models. The stream species quality component considered three metrics, the Index of Biotic Integrity for fish, aquatic macroinvertebrates, and mussels, which were indices generated by the Minnesota Department of Natural Resources and Minnesota Pollution Control Agency. The animal species richness component and at-risk animal species richness component were generated by using previous statewide survey data, where the mean watershed values for each species group were calculated and used as a comparison value to score every watershed (Minnesota DNR).

The connectivity sub-index included three components: terrestrial habitat connectivity, aquatic connectivity, and riparian connectivity. The terrestrial habitat connectivity component assessed the total habitat/connector area divided by the watershed, the patches of habitat and connections were generated by a habitat quality model. The aquatic connectivity component was derived by calculating the density of structures, including dams, culverts, and bridges, within the total length of the affected streams/rivers in the watershed. Riparian connectivity component was calculated by finding the percent agriculture and developed land area cover within 200 meters of the total riparian area (Minnesota DNR).

The geomorphology sub-index considered four components: soil erosion susceptibility, pollution sensitivity of near-surface material, climate water balance, and steep slope near streams. The soil erosion susceptibility component combined soil erodibility (K-factor) with a slope factor to produce a slope-scaled K-factor for each land area. The K-factors were averaged for each watershed to produce a linear scale for scoring. The pollution sensitivity of near-surface material component was calculated based on a pollution sensitivity model that was developed by the Minnesota Department of Natural Resource. The climate water balance component was produced by using the calculated water balance based on normal annual precipitation and evapotranspiration rates for Minnesota generated by local, quality-verified, weather sites (Minnesota DNR).

The hydrology sub-index had five components: perennial cover, impervious cover, water withdrawal, hydrologic storage, and flow variability. The percent perennial cover in the watershed was used directly as the perennial cover index score. Perennial cover was derived from a previous dataset from the Multi-Resolution Land Cover Consortium, and the land cover classes considered to be perennial were forest, shrub, grassland, pasture, and woody wetlands. The impervious cover index was calculated based on the total impervious area cover of the watershed. The imperviousness was determined by using satellite images, and the threshold for index score calculation was 4% based on previous studies. The water withdrawal index was generated based on the total permitted withdrawal from surface water and groundwater. The hydrologic storage index took two factors into account: water storage from lost wetland and water storage loss from altered watercourses. The water storage loss from pre-settlement to current was used to calculate an index score for hydrologic storage. The flow variability index was calculated using 33 Indicators of Hydrologic Alteration variables estimated from mean daily discharge data collected by USGS (Minnesota DNR).

The water quality component included three sub-categories: non-point source, localized pollution source, and assessments. The non-point source index was generated based on two metrics, the chemical application rates to agricultural lands and previous data on fertilizer and pesticides from the National Agricultural Statistic Service. The localized pollution source index took six point-sources into account: registered animal feedlots, potential contaminant sites, superfund sites, wastewater treatment plants, open pit mines, and septic systems. The assessment index was generated by calculating the total percentage of water systems that meet the EPA water quality standard for aquatic life, aquatic consumption, and aquatic recreation (Minnesota DNR).

The WHI of Minnesota was designed for simplified assessment of various ecological attributes of the watershed and for better understand the issues needing to be addressed. According to the Minnesota DNR, the purpose of developing the WHAF was to aid management decisions. The goal was to allow easier identification of factors that could change natural system response, which would be much quicker and less resource-consuming when directly addressing the main cause of the problem. The potential issue of the WHAF was that by taking the average of sub-index scores to produce a final index score, it assumed all the ecological indicators had equal influence on watershed health. This method ignored the possibility of ecological indicators having varied influence on watershed health and did not assign weighing factors for producing the final WHI score.

2.1.3 WHI review: Oregon State

In the Pacific Northwest, the city of Portland provided a watershed health index (WSHI) for five watersheds within the Portland city limits. It was described as a semi-quantitative tool to indicate condition changes in watersheds, and a translation of ecological data to information that can be more readily understood by the general public (City of Portland, 2019). The WSHI was also designed to highlight potential problematic sites and allow management teams to shift focus on sites that required the most attention. Another goal of the WSHI was to monitor watershed health changes over the short and long term. The index incorporated indicators from four types of ecological attributes: hydrology, water quality (water chemistry), physical habitat, and biological communities (City of Portland, 2019).

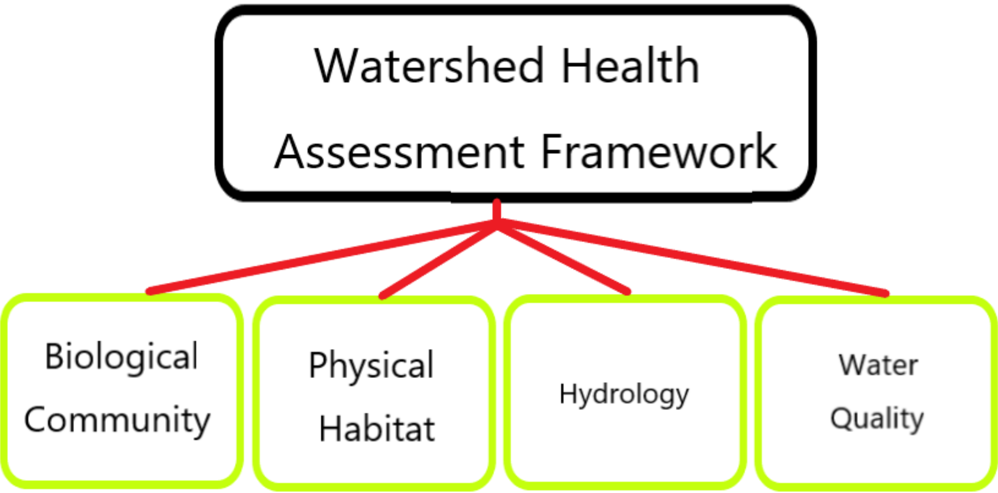


Figure 4. The conceptual figure of the main WHI structure in WSHI, the main WHI consists of 4 sub-indices with health scores that are aggregated to produce a health score for the main WHI.

Hydrological indicators factored in Effective Impervious Area (EIA) and Stream connectivity (% of piped streams), the purpose was to account for change in hydrology caused by man-made infrastructures and environmental alterations. Water indicators factor in nutrients ammonia-nitrogen, dissolved copper, dissolved oxygen, E. coli (individuals/100mL), water temperature, total mercury, total phosphorus, and total suspended solids. The water indicators could specify the potential source and type of pollution in the water systems. Physical habitat indicators factor in bank and flood plain conditions, riparian integrity, shallow water refugia and large woody debris count, substrate composition, and stream accessibility. Biological indicators factor in terrestrial communities (birds), aquatic communities (fish), and aquatic macroinvertebrate communities. Biological indicators were assessed with the Predictive Assessment Tool for Orgeon (PREDATOR), a predictive model that would be able to predict the occurrence probability of a taxa at a test site at a specific time (City of Portland, 2019).

The limitation for the WSHI scores generated for individual watersheds was that scores were not designed to be comparable between watersheds, but for monitoring change in condition over time for the same watershed (City of Portland, 2019). The WSHI was also limited to assessing streams and creeks, not larger rivers. The reason being that the data and model used for assessing larger waterbodies was different and current models may not be effective.

2.1.4. WHI review: State of California

The State of California developed an integrated watershed health assessment framework to identify and characterize the relative health of watersheds within the state (US EPA, 2012). Three major index categories were created to aggregate 23 indicators that characterize the health of watersheds: Watershed Condition, Stream Health, and Watershed Vulnerability (US EPA, 2012).

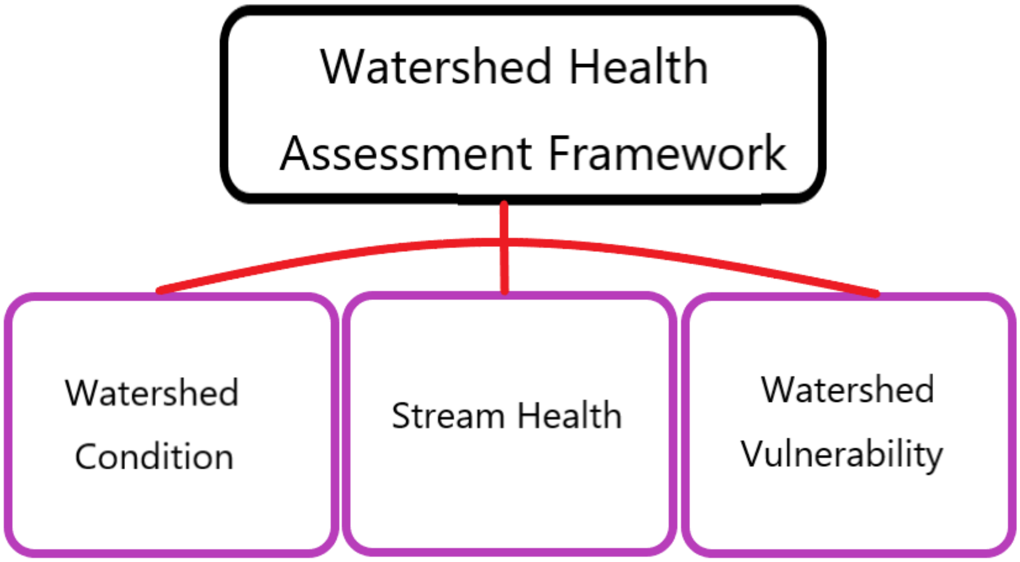


Figure 5. The conceptual figure of the main WHI structure in the integrated watershed health assessment framework, the main WHI consists of 3 sub-indices with health scores that are aggregated to produce a health score for the main WHI.

The watershed condition index included three sub-indices, relative watershed condition index, natural watershed condition index, and the anthropogenic watershed condition index. The three sub-indices accounted for percent natural land cover, percent intact active river area, sedimentation risk, percent artificial drainage area, dam storage ratio and road crossing density (US EPA, 2012). The stream health sub-index included four components: the relative stream health index, physical and biological habitat index, water quality index, and the instream biological condition index. Within the four components, indicators included were the California Rapid Assessment Method (CRAM) habitat assessment score, physical habitat assessment score, stream conductivity, stream nitrate concentration, stream turbidity, and California Stream Condition Index (CSCI) biological assessment score (US EPA, 2012). The watershed vulnerability sub-index encompassed five components: relative watershed vulnerability index, climate change vulnerability index, land cover vulnerability index, water use vulnerability index, and fire vulnerability index. The five sub-indices aggregated 11 indicators: projected change in precipitation, projected change in mean temperature, projected change in minimum temperature, projected change in maximum temperature, projected change in baseflow, projected change in snowpack, projected change in surface runoff, projected land cover change, water demand, projected change in wildfire severity, and fire regime condition class (US EPA, 2012).

2.2.1. Index structure and contents of WHI

While the U.S. EPA provided guidelines for the indicators to be included when developing a WHI, the examples from MN, OR and CA show that they are not strictly adhered to. The six ecological attributes mentioned in U.S. EPA's guidelines only provide a direction on what to monitor and did not include specific indicators to be included in the 6 ecological attributes; the decision was left to the state agencies to decide the ecological indicators to be included in the WHI.

Table 1. Summary of the sub-indices in the WHI of the three states (MN, OR, and CA).

WHAF - MN	Biology	Connectivity	Geomorphology	Hydrology	Water Quality
WSHI - OR	Biological Community	Physical Habitat	Hydrology	Water Quality	
IWAF - CA	Watershed Condition	Stream Health	Watershed Vulnerability		

The structures of WHIs reviewed indicated that the main index structure includes multiple sub-indices, with each of the sub-indices representing one type of ecological attribute or water quality factor. The common approach is to aggregate individual health indicators into sub-indices by compiling the indicators using conversion factors determined from past studies (Minnesota DNR), then sub-indices are manipulated to create the comprehensive index score. The purpose of using sub-indices to represent ecological attributes or water quality factor is to gather all possible information from every aspect of the watershed to create a comprehensive health assessment of the watershed.

The WHIs in Table 1 all appear to include indicators within three categories: physical habitat, biological condition, and water quality. Indicators of physical habitat is known to influence the living conditions of organisms within the watershed, and poor physical habitat conditions can result in reduced ecosystem services (Maddock, 2001). Biological indicators are commonly used as a measurement of environmental degradation and human impact, as some stream organisms are sensitive to changes in the surrounding environment (Tanaka et al., 2016). Water quality indicators can describe the overall condition of water systems within the watersheds and can be used to specify the impact of human activity in the surroundings (Mimikou et al., 2000).

The common indicators selected for biological health is the benthic community structure, which indicates the significance of benthic community structure on watershed health (Maddock, 2001). The benthic community structures are measured with Benthic Index of Biological Integrity (B-IBI), with each state using different B-IBI calibrated for the local environment. Area calibrated B-IBI better represents the biological health of stream systems and allow researchers to standardize assessments within the specified region (WA DNRP, 2014). Common water quality indicators included in the three WHIs measure water conditions and chemical concentrations, including conductivity, dissolved oxygen level, pH, water temperature, Phosphorus concentration, and Nitrogen concentration. Indicators such as, conductivity, dissolved oxygen level, water temperature, and pH, have heavy influences on the health of biota in the water, which is crucial for watershed health (Tanaka et al., 2016). Indicators that represent chemical concentrations, such as P concentration and N concentration, can describe anthropogenic input of nutrients and help with watershed management effort (Hascic and Wu, 2006).

2.2.2. Potential disadvantages of previous WHIs

WHI of Minnesota state and the City of Portland compiles the final WHI by taking the average of sub-index scores, which assumes all sub-indices have the same weight factor that makes up the final WHI score. The potential issue with such a method is that it does not account for potential differences in contributions different environmental indicators may have on watershed health. The more ideal approach is to create sub-indices and main index weight factors through the analysis of local watershed health data, which allows the data to provide unbiased objective information on the contribution of different watershed health indicators on the health of the watershed.

The watershed health indicators selected for watershed health assessment in the three WHIs varies from continuous numerical data to categorical data collected through human judgement. While numerical data is a direct measurement of an indicator, categorical data collected through human judgement can result in biased data that may not best represent the actual conditions of the watersheds. To increase the accuracy of WHI, it is best to avoid human subjectivity by using a data-driven approach.

Chapter 3: Development of a Watershed Health Index for Washington State

3.1. Introduction

3.1.1 A Data Driven WHI for Washington State

The health condition of watersheds is a critical factor in a properly functioning ecosystem, both natural and anthropogenic activities heavily depend on a functioning watershed to provide ecosystem services (Hazbavi et al., 2016). Lack of understanding and care on watershed health can result in both environmental disaster and socioeconomical issues, which is why watershed health, known as the holistic condition of freshwater ecosystems within a watershed, is critical in maintaining the delicate balance of nature (Mirchi et al. 2009; U.S. EPA, 2012). Watershed characteristics often vary widely depending on the physical location of the sites, making it difficult to create generalized evaluation methods for watershed health for the entire United States (U.S. EPA, 2013).

Watershed health monitoring has been an ongoing effort in Washington State, but it currently does not have a watershed health index. However, Washington State does have a watershed health monitoring program that aims to track trends and consistently monitor watershed health conditions. The Washington State Department of Ecology (WA DoE) recently conducted a statewide stream macroinvertebrate bioassessment to assess the environmental stressors that influence the biological integrity of the stream systems and found 60% of stream stretches to be in poor biological condition (Larson et al., 2019). The benthic macroinvertebrate community integrity is commonly used as an indicator for stream health (Wetzel, 2001), thus the statewide stream macroinvertebrate assessment results are concerning and have implications for declining freshwater quality and the health of the watersheds in the state.

The WA state agencies have put in effort to better understand watershed health and freshwater systems. The statewide bioassessment assessed stream health by associating risk factors of health indicators to the Benthic Index of Biotic Integrity, or B-IBI (Larson et al., 2019), which solely assessed the biological health of the stream systems and not all aspects of watershed health. While the B-IBI is a good indicator for stream health, watershed health assessments need the inclusion of more environmental attributes, such as physical habitat and water chemistry, because of the complexity of freshwater systems. The database from Washington State's watershed health monitoring effort includes data from stream biological, physical and chemical attributes, thus making it possible to assess watershed health with appropriate methods.

To develop a WHI for Washington State, past monitoring data would be required for analysis to better understand the relationship between the ecological indicators and watershed health. WA DoE allocated resources to monitor watersheds within the state and created a database with watershed data for approximately a decade. The data was previously used to conduct a statewide stream macroinvertebrate bioassessment but not actively used for developing a watershed health index at this point. The goal of this study is to use the available data from WA DoE's database to develop a WHI for Washington State. Development of a WHI for Washington state using B-IBI and historical water quality data provides an opportunity to apply newly developed data science tools.

To utilize new data science tools for our WHI, we will need understand the general structure for constructing a WHI. Developing a watershed health index would require the use of an existing stream health indicator as a reference for the health of the stream systems. One of the indicators most used for stream health nationwide would be the assessment of benthic macroinvertebrate communities in the streams (Wetzel, 2001). For Washington State, the Benthic Index of Biological Integrity (B-IBI) is a multi-metric index that is commonly used to assess the benthic macroinvertebrate communities (Larson et at., 2019). The B-IBI index is made up of 10 metrics: Total taxa richness, Ephemeroptera taxa richness, Plecoptera taxa richness, Trichoptera taxa richness, long-lived taxa, intolerant taxa, percent tolerant individual, clinger taxa, percent predator individuals, and percent dominance.

Finally, we need to use modern tools to analyze the data. When analyzing a large dataset with multiple variables, dimensionality may be an issue when trying to understand relationships between variables. To reduce dimensionality of the dataset while minimizing loss of information, unsupervised machine learning, such as principal component analysis (PCA), can be used to understand water quality data (Karim and Taha, 2003). Unsupervised learning, such as PCA, aims to infer the original structure of the dataset; this allows for the understanding of relationships between variables without prior knowledge of the outcome. Another approach is using Supervised learning, such as Partial least square regression (PLS), are analytic tools that is capable of incorporating large datasets when understanding relationships in water quality data. PLS is a multivariate regression analysis that overcomes the problem of dimensionality in large multi-variable datasets (Khatri et al., 2020). Supervised learning requires prior knowledge of the outcome, in this case a known stream health index, to estimate the relationship between predictor variables and the outcome. Since this study would be using Benthic Index of Biological Integrety (B-IBI) as a reference for stream health condition, using supervised learning methods, such as PLS, would show how well the response variables are predicted by the predictor variables. Another supervised learning method used in watershed health and ecology studies is the Boosted Regression Tree

(BRT), as it has multiple advantages over linear regression (U.S. EPA, 2012). The BRT is capable of predicting most types of data, including numerical and categorical; BRT is insensitive to potential outliers in the dataset, allowing for easy data processing; BRT does not assume linear relationships, and capable of accounting for non-linear relationships between predictor and response variables.

The goal of this study is to create a WHI for Washington State that can assess watershed health with environmental indicators that are measurable, while including indicators that are not used in other health indices such as biological integrity indices and water quality indices. This would allow for parallel assessments of watershed health with different indices and help better understand the health condition of the watersheds. Unlike past examples where sub-indices included subjectively determined weighting factors, we aim to create an objective method of determining weighting factors using the available datasets and analytic tools. Eliminating subjective decisions can minimize human bias and allow the data to express the relationships between variables.

3.2. Materials and Methods

WA DoE's watershed monitoring effort included collecting stream and river samples from 8 regions in the state since 2009. The monitoring objective was to characterize the biological, chemical and habitat condition of the streams and rivers, which would project an overall health condition of watersheds. The data collection was completed within a day, meaning that the data was a "snapshot" of the condition of the watersheds, and not continuously monitored data. The data can be divided into 3 attributes: biological, chemical, and physical habitat.

Biological habitat included benthic macroinvertebrate counts and vertebrate counts. Benthic macroinvertebrates were identified with taxon levels (family, genus, species) and life stage (larva, adult). Vertebrates were identified with scientific names, common names, and life stages (juvenile, adult, unknown) as well.

Water chemistry included conductivity, pH, dissolved oxygen amount and percentage, water temperature, turbidity. Water chemistry was measured once early in the day and once more later in the day for every site. Sediment chemistry included inorganic (Cu, Pb, Zn, N, P) and organic compound concentrations.

Physical habitat included large woody debris, human influence, riparian zones, substrate composition and fish cover. Large woody debris counts were categorized by diameter and length. Human influence data identified land use (buildings, clearing, paths, landfills, mining activities, park/lawn, paved roadway/railway, pipes, trails, and wall/dam) near the sample site. Riparian zone data identified vegetation coverage of canopy, understory, and ground cover. Fish cover data recorded the percentage water surface coverage of different cover type (artificial structures, boulders, woody debris, bryophytes, filamentous algae, live trees, macrophytes, overhanging vegetation, undercut banks).

The database included 1467 data collection entries, with around 500 individual sites that were visited one to three times throughout the ten or more years. Approximately 460 entries had incomplete datasets, 25 entries had no benthic macroinvertebrate data and 72 entries only included benthic macroinvertebrate data. The data collection was a joint effort between seven agencies and organizations, and a Standard Operating Procedure (SOP) was followed during all the data collection events. However, selection of data was decided by individual agencies and organizations, thus some collection events did not include all data assigned in the SOP. Selection of data was based on the ability of the entity to process samples and the availability of resources.

3.2.1. Indicator selection

Development of a watershed health index requires selection of environmental indicators for data exploration and analysis. The U.S. Environmental Protection Agency (U.S. EPA) guidelines suggest indicators from six essential ecological attributes for watershed health assessment: landscape condition, habitat, hydrology, geomorphology, water quality, and biological condition (U.S. EPA, 2012). However, the U.S. EPA does not specify specific indicators to be included in a WHI, and to include all possible environmental indicators within the six ecological attributes requires large monitoring efforts that may not be achievable depending on budget and resource availability. For example, WHIs developed by the City of Portland and the state of Minnesota each selected environmental indicators based on the need determined by the local agencies, and not strictly by the U.S. EPA's recommendation. In this study, we focused on selecting environmental indicators available in the WHMweb database held by the Washington State Department of Ecology. We selected indicators that consisted continuous numerical variables and avoided categorical data. Extra environmental variables, annual precipitation and annual traffic count, were added through the use of ArcGIS.

3.2.2 Data Processing

Since this study aims to use a data-driven approach to create a WHI that could best describe the relationship between watershed health and ecological indicators, we need to eliminate ecological data that was recorded based on human judgement. Upon examination of the data collection SOP for the dataset provided by Washington Department of Ecology, we noticed that the collection of categorical data requires the data collectors to make judgement calls based on observations. Before proceeding to the analysis, modification to the dataset is needed to only include direct measurements that did not involve any human-based judgements, such as the chemical concentrations in sampled streams.

Continuous numerical environmental indicator data was extracted from the WHMweb for processing, categorical data was not included in the extracted dataset to minimize human bias. A major limitation in the WHMweb data is that not all indicators were sampled for every data collection event, resulting in partially missing data within the dataset. Some multivariate analysis models, such as Partial Least Square regression, requires the input of a complete dataset and missing data was not acceptable as it would not allow proper functioning of the models (Wan et al.. 2020). Imputed averages were used to fill in missing values thereby addressing the issue with missing data. Imputed averages should not influence the overall distribution of the data and allows for the use of multivariate analysis on the dataset.

3.2.3 Data Exploration:

The dataset includes over 1000 entries and 24 categories, which has high dimensionality and requires dimension reduction to understand. This study uses machine learning to perform dimension reduction techniques to find relationships between watershed health and the health indicators. We used an unsupervised learning approach to assess initial relationship by running Principal Component Analysis (PCA), a routine used to reduce the dataset into a two-dimensional plot. The PCA provides loadings and a biplot that allows visualization of the relationships of health indicators while minimizing bias since PCA is an unsupervised learning method.

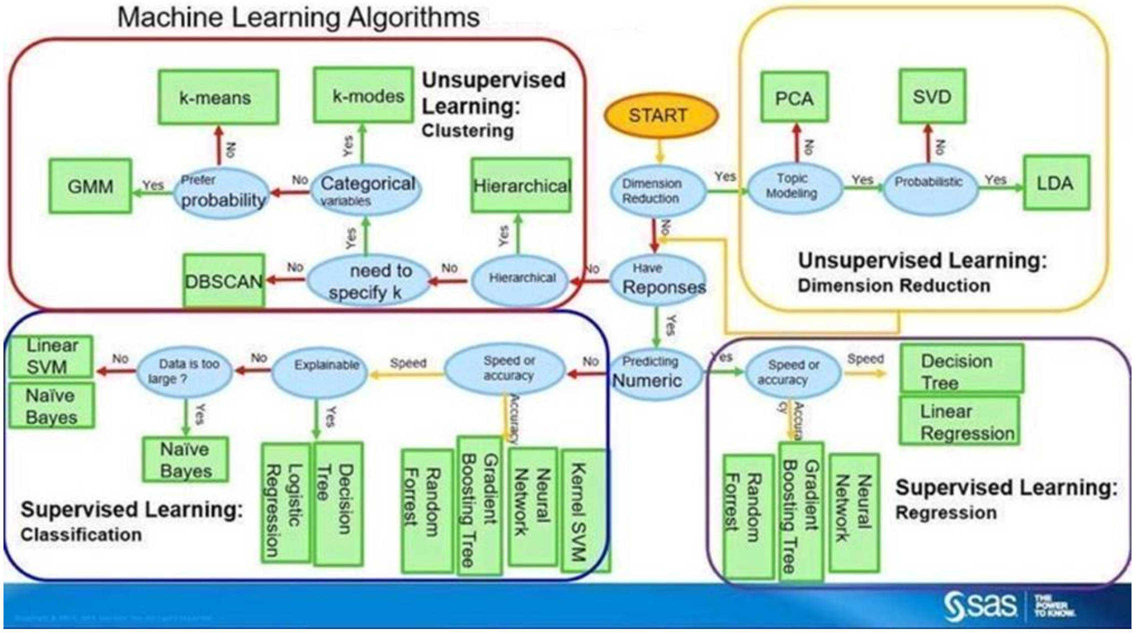


Figure 6. Machine learning logic tree that guides researchers to select appropriate analysis methods during data processing (Hui, 2020)

To further explore the dataset, we need to maintain the dimensionality of the dataset to avoid loss of information, we also need to assign a response variable to associate the stream health indicators to watershed health. According to Figure 6, datasets with numeric variables with response variables require the use of supervised learning, such as Partial Least Square Regression (PLS) and Boosted Regression Tree (BRT). PLS is known as a robust tool for multivariable statistical data analysis in water quality studies, isolating issues such as collinearity and autocorrelation (Khatri et al., 2020). BRT is used in previous construction of watershed health assessment frameworks in California, it can process all data types and it is insensitive to outliers or incomplete datasets (U.S. EPA, 2012).

We included more watershed health indicators into the construction of our WHI, data from external sources was included in the analysis. Two indicators, annual precipitation and traffic count, were extracted from ArcGIS using the maps provided by Washington State Department of Transportation. GPS coordinates from the DoE dataset were extracted and imported into ArcGIS to locate all sampled sites, data maps were laid over and additional data was obtained by interpolating between available data points on the GIS maps.

Principal Component Analysis

We used python to standardize the dataset and input the dataset into the Principal Component Analysis model. The input variables were total taxa richness , Ephemeroptera taxa richness, Plecoptera taxa richness, Trichoptera taxa richness, long-lived taxa, intolerant taxa, percent tolerant individual, clinger taxa, percent predator individuals, and percent dominance, conductivity (µS/cm), total P (µg/L), water temperature (°C), dissolved O_2 (mg/L), suspended solid (mg/L), pH, total N (µg/L), Flow (cfs), percent fine sediments (%), coordinate index calculated by multiplying Latitude by Longitude that can represent individual locations with a unique numerical value, annual traffic count (passes/yr), annual precipitation (in).

3.2.4 Designing the general structure of the WHI

The U.S. Environmental Protection Agency (US EPA) suggests creating a WHI structure consisting of sub-indices that describe different ecological attributes (U.S. EPA, 2012). Existing WHI structures in other states, such as Minnesota State, linked corresponding environmental indicators to the individual sub-indices to output sub-index scores, and the sub-index scores are manipulated by a pre-determined equation to calculate the final WHI index score (WA DNRP, 2014). We will be creating our WHI based on a similar structure suggested by the U.S. EPA, with one main index comprising of multiple sub-indices that represent different aspect of stream health and linking watershed health indicators to each sub-indices. The conceptual structure of the WHI is shown in Figure 7. The main Watershed Health Index score is a compilation of multiple sub-index scores, and every sub-index is a compilation score of individual watershed health indicator measurements evaluated through previous studies. We will use analytic tools to determine the influence of watershed health indicator measurements on the variability of sub-index scores, which will determine the relative influence of each watershed health indicator on the variability of each sub-index.

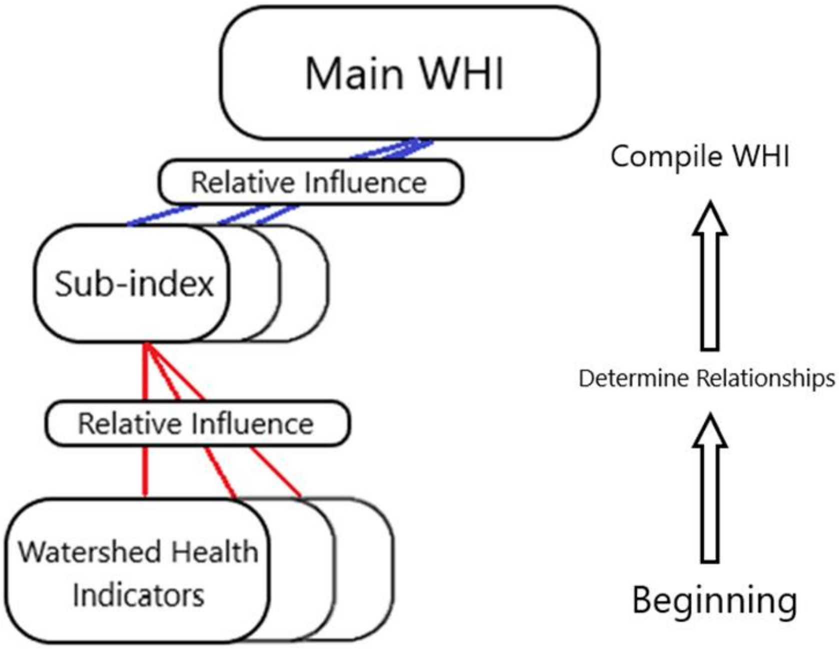


Figure 7. Conceptual figure of the general structure of the WHI. The main WHI consists of a summarized score of the multiple sub-indices. The relative influences of watershed health indicators on the variability of sub-indices are determined with analytic tools. Similarly, the relative influences of sub-indices on the variability of the main index are also determined with the same analytic tools.

We used the Benthic Index of Biointegrity (B-IBI) as the health metric for measuring the health of streams sampled in the dataset. The B-IBI contained ten metrics, each metric describes the health condition of a specific group of benthic macroinvertebrates that can indicate different types of human disturbance (Wetzel, 2001). The ten metrics in B-IBI were each used as a sub-index during the WHI development since each metric included a portion of the information on the biological health of the evaluated stream (WA DNRP, 2014). The B-IBI is used as the watershed health metric for developing our WHI, which dictates that our WHI will have a similar structure as the B-IBI. The final WHI comprises ten sub-indices that correspond to the ten metrics within the B-IBI. To determine the correlation between the B-IBI and the ten metrics, the B-IBI was disbanded into the ten metrics and investigated. The conceptual pipeline is displayed in Figure 8. The relative influence of the B-IBI metrics on B-IBI score and the relative influence of watershed health indicators on individual B-IBI metrics were determined through the use of supervised learning methods, Partial Least Square regression and Boosted Regression Tree. The relative influences are then used to convert individual watershed health indicators into a main index score.

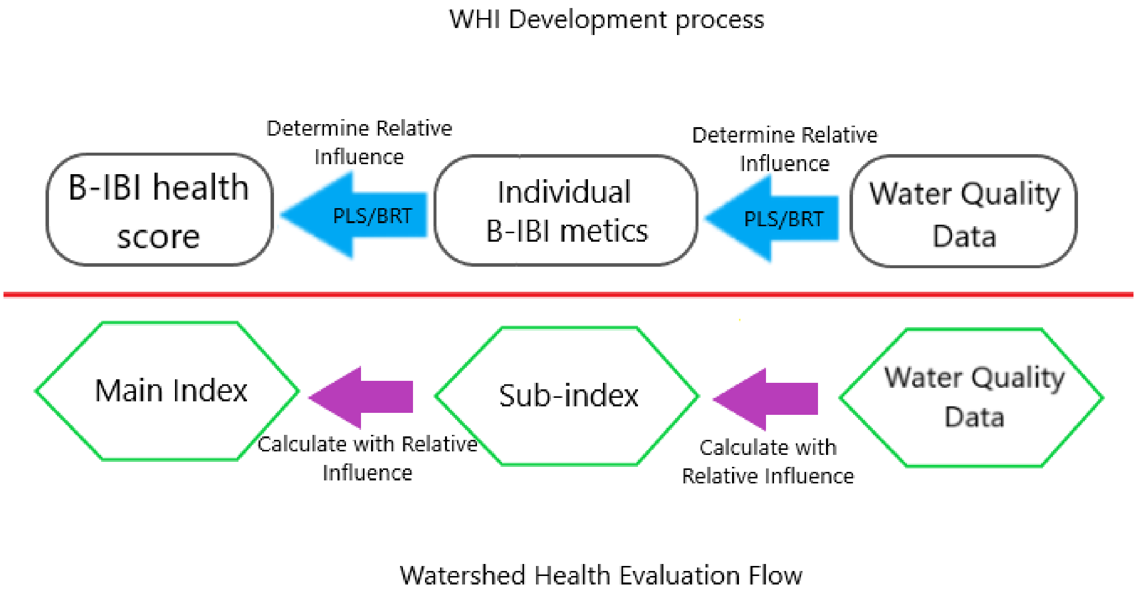

WHI Development process
Determine Relative Influence
B-IBI health score
PLS/BRT
Individual B-IBI metics
Determine Relative Influence
PLS/BRT
Water Quality Data
Main Index
Calculate with Relative Influence
Sub-index
Calculate with Relative Influence
Water Quality Data
Watershed Health Evaluation Flow

Figure 8. Conceptual pipeline of creating and using our WHI. The correlation between B-IBI score and individual metrics, and the correlation between individual metrics and watershed health indicators are determined by PLS and BRT analysis. Main WHI score is calculated by using relative influences as conversion factors between indicators and the main index.

3.2.5 Creating the WHI structure with Boosted Regression Tree

To determine the relationship between the B-IBI and its ten individual metrics, the B-IBI was entered as the response variable and the ten individual metrics were entered as the predictor variables in the Boosted Regression Tree (BRT) model. The BRT model output reports the relative influence of the ten individual metrics on B-IBI, with a maximum possible relative influence of 100%. The relative influences of the ten metrics were used as the weight factors for the ten corresponding sub-indices that made up the final WHI. Linking the stream health indicators to the sub-indices was the next crucial step.

The stream health indicators were entered into the BRT model as the predictor variables and each of the ten B-IBI metrics were entered as the response variable separately, which resulted in ten BRT model results. The relative influences of the stream health indicators on the ten individual B-IBI metrics from the BRT models were used as weight factors to convert the stream health indicators into the ten sub-index scores. The WHI structure created with BRT present a percentage relationship between the individual watershed health indicators and final WHI score

3.2.6 Creating WHI structure with Partial Least Square regression

PLS regressions was performed with B-IBI as the response variable, and each of the ten B-IBI metrics as the predictor variables, which resulted in ten PLS regressions. The explained variances of variables by PLS-components from each of the ten PLS regressions were generated and were summed up to make up a total explained variance for the ten PLS regressions. The individual explained varianc es of each of the ten regressions were divided by the total explained variance to generate the percent influence of each metric on the B-IBI. The calculated percent influence was used as the weight factors for the WHI sub-indices to calculate the final WHI, and the main structure of the final WHI consisted of 10 sub-indices that added up to 1.

The next step was to correlate the stream health indicators with the 10 individual B-IBI metrics, which corresponded to the ten sub-indices. To determine the relationship between each B-IBI metric and the stream health indicators, PLS regressions were performed. Each of the B-IBI metrics were used, individually in separate PLS regressions, as the response variables and the stream health indicators were used as the predictor variables, which resulted in ten PLS regressions. Similar to the generation of the main WHI structure, the explained variance of variables by PLS components were summed up to a total explained variance, and the explained variance of each stream health indicators was divided by the total explained variance, to generate the percent influence of each stream health indicators on the sub-indices. The calculated percent influence was used as weights to convert the stream health indicators into the final sub-indices scores. The WHI structure created with PLS present a percentage relationship between the individual watershed health indicators and final WHI score

3.2.7. Conversion of sample data to health scores

Sample data of the individual watershed health indicators will need an initial assessment to be converted into indicator health scores before being entered into the WHI assessment. To create the initial assessment system, we will isolate sample sites with B-IBI higher than 75% from the full dataset as the healthy group. Other sample sites with B-IBI lower than 75% are considered not healthy enough and will be excluded for this section. The means and standard deviations of the individual watershed health indicator are calculated. The calculated means of the watershed health indicators in the healthy group are used as the 100% thresholds for converting watershed health indicator data into indicator health scores, and deviations from the 100% thresholds of 2 standard deviation will result in scores of 0%. The means and standard deviations of watershed health indicators in the healthy group will be used to linearly scale the initial assessment scoring system as shown in Figure 9. The threshold for good condition will be at 50% and higher, or 1 standard deviation or less from the mean. Any indicator measurement that deviates more than 2 standard deviations from the mean will be considered in poor condition and receive a score of 0%.

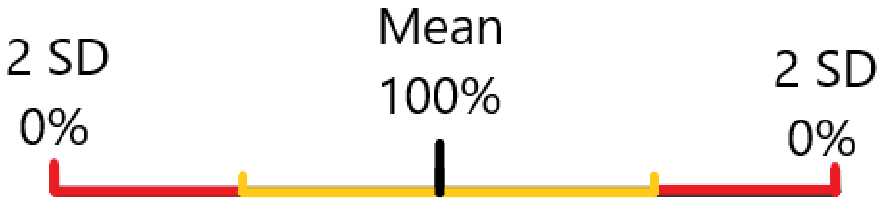


Figure 9. Illustration of the linear scoring scale of the initial conversion of watershed health indicator data to indicator health score.

3.3. Results

3.3.1 Unsupervised Learning: Principal Component Analysis

Results from PCA indicated that Principal Component 1 (PC1) and Principal Component 2 (PC2) explained 27.47% and 9.67% of the variance in the dataset, respectively.

Loadings of each stream health indicator on PC1 and PC2 is presented in Table 2, the PCA plot and graphical presentation of the loadings are presented in Figure 10. There was no major clustering identified in the PCA plot, samples were scattered across the plot.

Table 2. PCA Loadings of each stream health indicator variable on PC1 and PC2.

Variable	**Principal Component 1**	**Principal Component 2**
Total Taxa Richness (total)	-0.5885	-0.3425
Ephemeroptera Richness (may)	-0.7523	0.0220
Plecoptera Richness (stone)	-0.7896	-0.0115
Caddisfly Richness (caddis)	-0.7012	-0.3467
Long-lived Richness (long)	-0.6903	-0.3100
Intolerant Richness (intolerant)	-0.8004	-0.0023
Clinger Richness (clinger)	-0.7486	-0.4200
Predator Percent (perc_pred)	-0.1973	0.1343
Tolerant Percent (perc_tol)	0.3430	-0.0377
Percent Dominant (perc_dom)	0.4805	0.2269
Conductivity	0.5426	-0.3772
Total Phosphorus (total P)	0.4460	-0.1972
Water Temperature (water temp)	0.4431	-0.2606
Dissolved Oxygen	-0.3925	-0.0477
Suspended Solids	0.2025	0.1198
pH	0.1755	-0.5898
Total Nitrogen	0.4010	-0.2423
Stream Flow (flow)	0.0361	-0.0693
Percent Fine Sediments (% fine)	0.3315	-0.3102
Traffic Count	0.805	0.1441
Annual Precipitation	-0.3137	0.6557
Lattitude x Longitude	-0.2182	0.6174

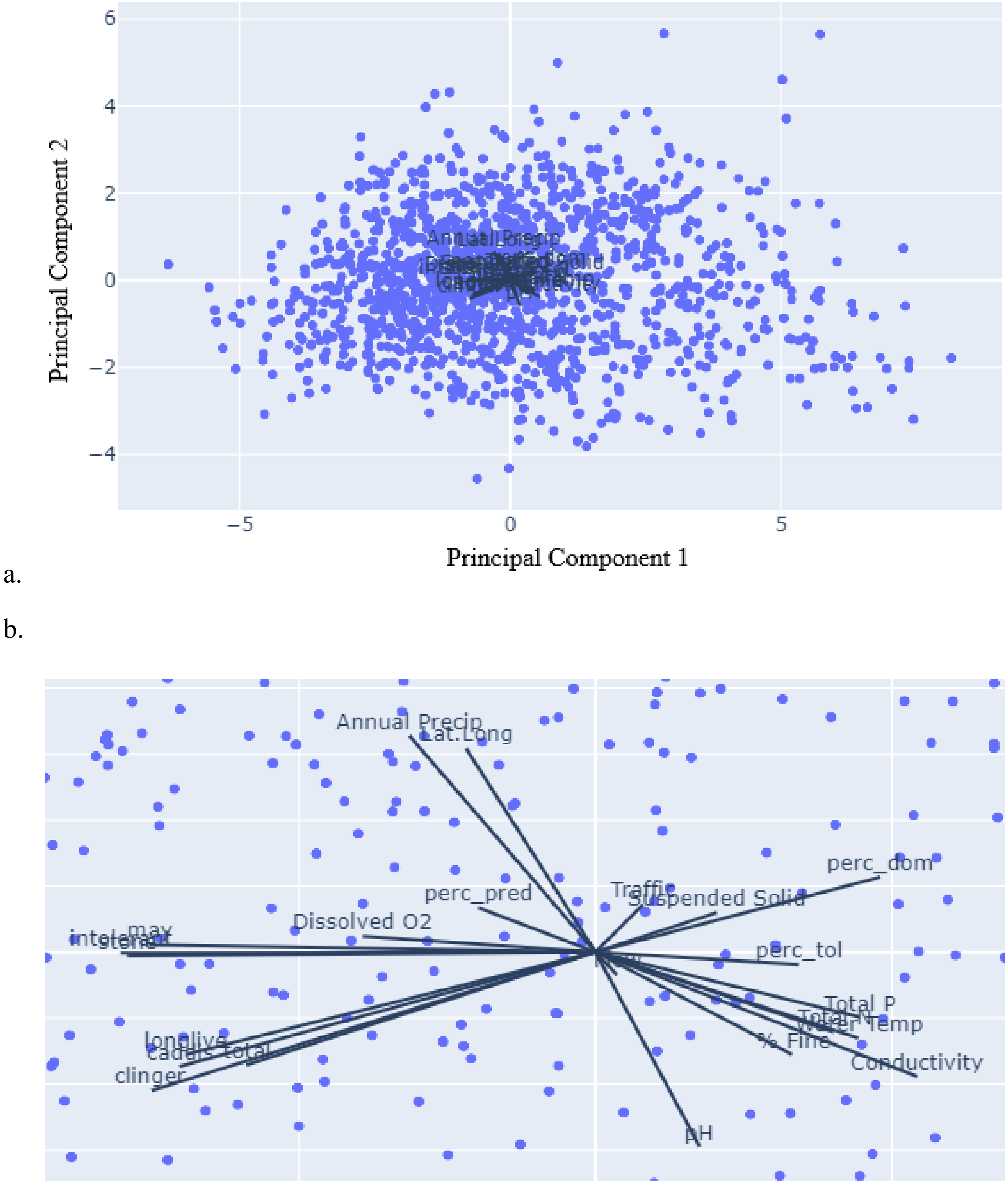


Figure 10. (a) PCA biplot, with principal component 1 on the x axis and principal component 2 on the y axis. (b) Loadings plot of each stream health indicator variables, as presented in

table 1.

The B-IBI metrics showed strong correlation with dissolved oxygen level and suspended solids, other stream health indicators did not show a strong correlation with the B-IBI metrics, especially for flow and pH. One of the B-IBI metric, Percent predator, showed a weak correlation with most stream health indicators. Annual precipitation and location (Lat.Long) showed strong correlation with most of the stream health indicators, with the exception of dissolved oxygen level and suspended solids. Flow had very minimal impact on all other variables.

3.3.2 Supervised Learning: Boosted Regression Tree

BRT results of using B-IBI as response variable and the ten B-IBI metrics as predictor variables is shown on Figure 11. The relative influence of intolerant taxa was the highest at 33.52% and the lowest relative influence was the percent tolerant individual at 2.12%. Clinger taxa had the second highest relative influence at 14.82% and the relative influence of other metrics were all below 10%.

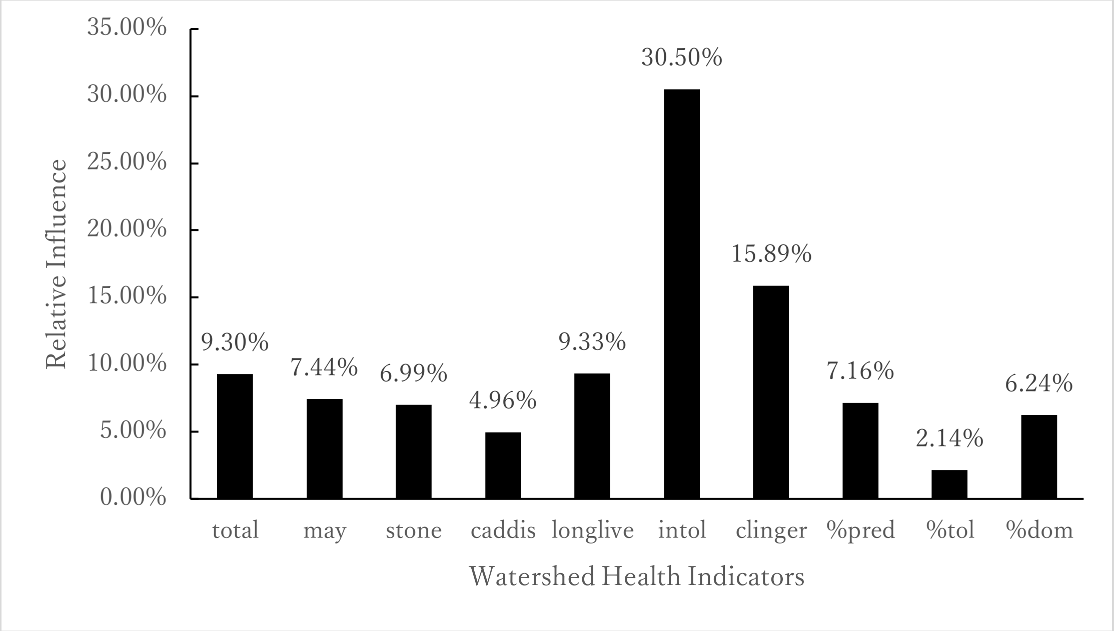


Figure 11. Result of Boosted Regression Tree (BRT) expressing the relative influence of individual B-IBI metrics on B-IBI in percentage, with all relative influence adding up to 100%.

The second part of the BRT analysis consisted of ten BRT results using the individual B-IBI metrics as response variables and the stream health indicators as the predictor variables. The relative influences of location of sample site (Lat.Long) and suspended solids on Total taxa richness were the highest at 13.56% and 13.22%, respectively, and the lowest relative influence was annual precipitation at 3.25%. Conductivity had the second highest relative influence at 12.88% and the relative influence of other stream health indicators were all below 10%. The relative influences of water temperature on Ephemeroptera richness was the highest at 30.01%, and the lowest relative influence was suspended solids at 2.15%. Location (Lat.Long) had the second highest relative influence at 12.48% but much lower than that of water temperature. The relative influence of conductivity had the most impact on Plecoptera taxa richness at 18.19%, water temperature was close behind with a relative influence of 15.70%; annual precipitation was the lowest at 1.26% relative influence. The highest influence on Trichoptera taxa richness was dissolved oxygen level and conductivity, at 13.47% and 12.39% respectively; annual precipitation had the lowest relative influence at 2.35%. Long-lived taxa seemed to be influenced by location (13.68%), flow (12.03%) and conductivity (11.12%) the most and least influenced by annual precipitation at 1.71%. The intolerant taxa was most influenced by water temperature with a relative influence of 20.54% and least influenced by annual precipitation at 2.04%. Stream flow had the most influence on clinger taxa richness at 15.67% and annual precipitation had the least influence on clinger taxa richness at 2.19%. Percent predator individuals was most influenced by stream flow at 19.73% and least influenced by annual precipitation at 2.14%. Conductivity had the highest relative influence on percent tolerant individual at 18.60% and annual precipitation had the least influence at 1.19% (Figure 6). Percent dominance was most influenced by location, stream flow and dissolved oxygen level at 13.32%, 11.25% and 11.21% respectively; annual precipitation had the least influence on percent dominance at 1.86% (Figure 12).

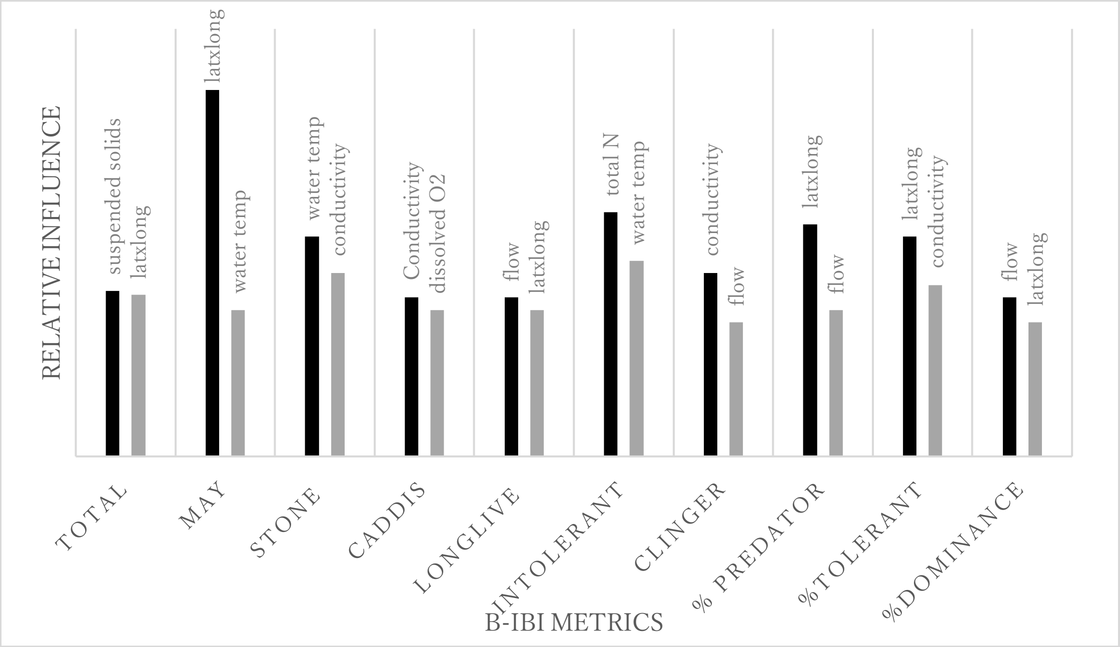


Figure 12. Result of Boosted Regression Tree (BRT) expressing the relative influence of stream health indicators on individual B-IBI metrics in percentage, with all relative influence adding up to 100%. Each of the two bars represented the top two relative influence of stream health indicator on each B-IBI metric of Total taxa richness, Ephemeroptera taxa richness, Plecoptera taxa richness, Trichoptera taxa richness, long-lived taxa, intolerant taxa, clinger taxa richness, percent predator individuals, percent tolerant individual, and percent dominance. Bottom bar of each category represents the highest relative influence and the top bar represents the second highest relative influence.

There were five indicators that frequently showed high influence, above 10% relative influence, on the ten B-IBI metrics: location, stream flow, dissolved oxygen level, conductivity and water temperature. Annual precipitation had the least relative influence on most B-IBI metrics, with relative influence percentage frequently below 3%.

3.3 Supervised Learning: Partial Least Square Regression

The first part of the Partial Least Square regression analysis involved using B-IBI metrics as predictor variables and B-IBI as the response variable. Percent predator and clinger taxa richness had the most influence on B-IBI at 13.47% and 12.87% respectively, and percent tolerant individual had the lowest relative influence at 4.7% (Figure 13). Overall, the relative influence of the B-IBI metrics were mostly even at around 10% with the exception of the aformentioned metrics showing more extreme percentages.

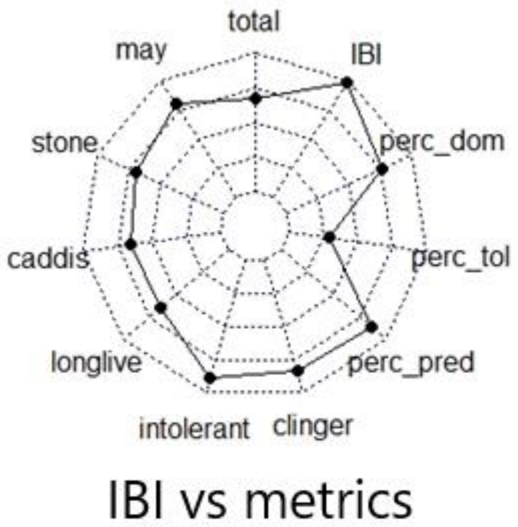
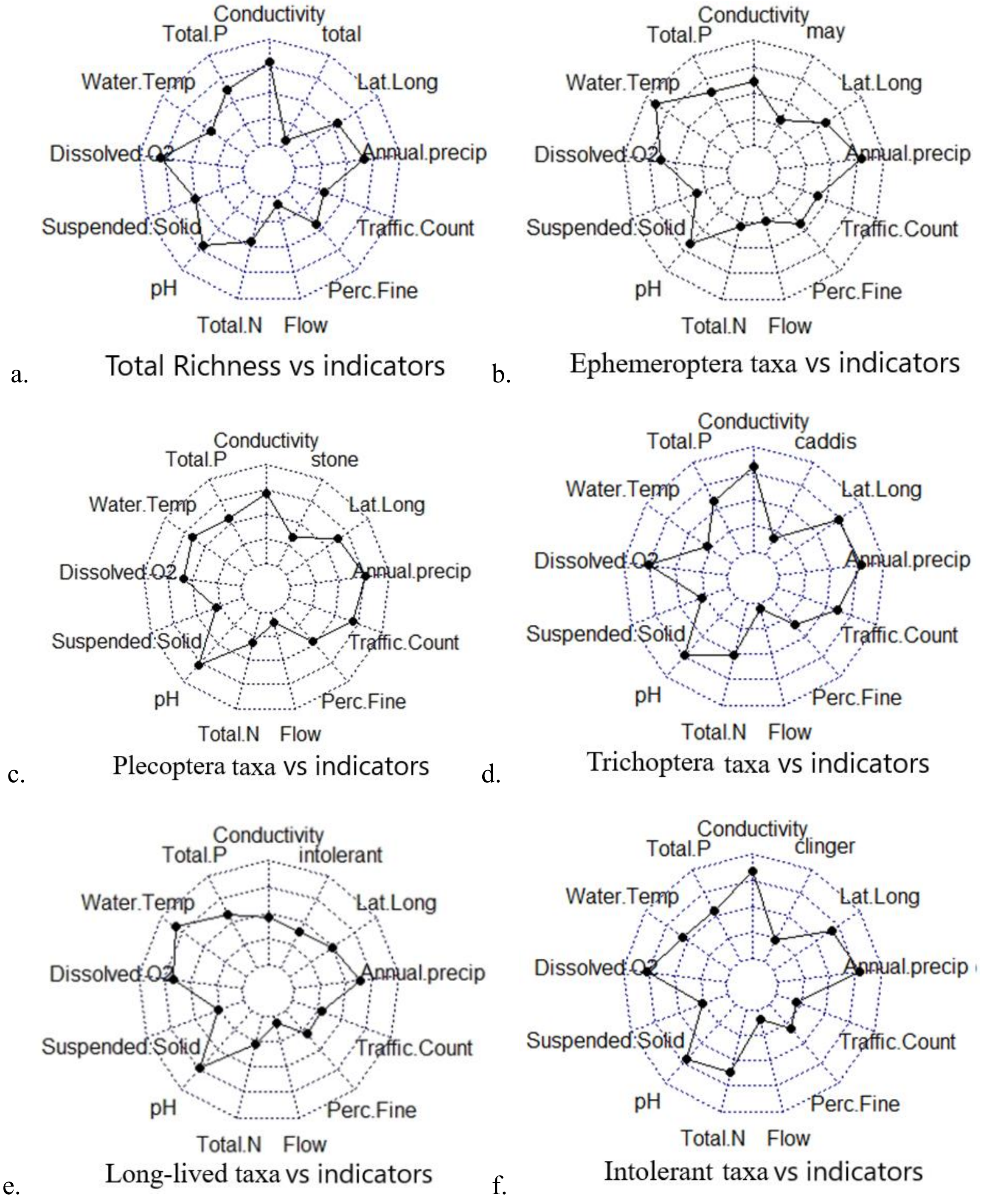


Figure 13. Result of Partial Least Square regression (PLS) expressing the relative influence of individual B-IBI metrics on B-IBI in percentage, with all relative influence adding up to 100%.

The second part of the PLS analysis used watershed health indicators as predictor variables and the ten individual B-IBI metrics as the response variables. In Figure 14(a), the relative influences of dissoveled oxygen level and conductivity on Total taxa richness were the highest at 12.50% and 12.57%, respectively, and the lowest relative influence was stream flow at 1.18%. The relative influences of water temperature and annual precipitation on Ephemeroptera richness was the highest at 13.74% and 12.14%, and the lowest relative influence was stream flow at 3.71% (Figure 14(b)). The relative influence of water pH had the most impact on Plecoptera taxa richness at 12.00%, and annual precipitation was close behind with a relative influence of 11.64%; stream flow was the lowest at 1.62% relative influence (Figure 14(c)). The highest influence on Trichoptera taxa richness was conductivity at 12.49%, and stream flow had the lowest relative influence at 0.75% (Figure 14(d)). Long-lived taxa seemed to be influenced by annual precipitation (13.07%) the most and least influenced by stream flow at 5.20% (Figure 14(e)). The intolerant taxa was most influenced by water temperature 13.43% and least influenced by stream flow at 2.36% (Figure 14(f)). Conductivity had the most influence on clinger taxa richness at 13.18% and stream flow had the least influence on clinger taxa richness at 1.51% (Figure 14(g)). Percent predator individuals was most influenced by water temperature at 14.80% and least influenced by traffic count and stream flow at 3.81% and 3.52% respectively (Figure 14(h)). Location had the highest relative influence on percent tolerant individual at 13.62% and suspended solids had the least influence at 2.45% (Figure 14(i)). Percent dominance was most influenced by conductivity at 13.00%, and stream flow had the least influence on percent dominance at 0.52% (Figure 14(j)).



a. Total Richness vs indicators

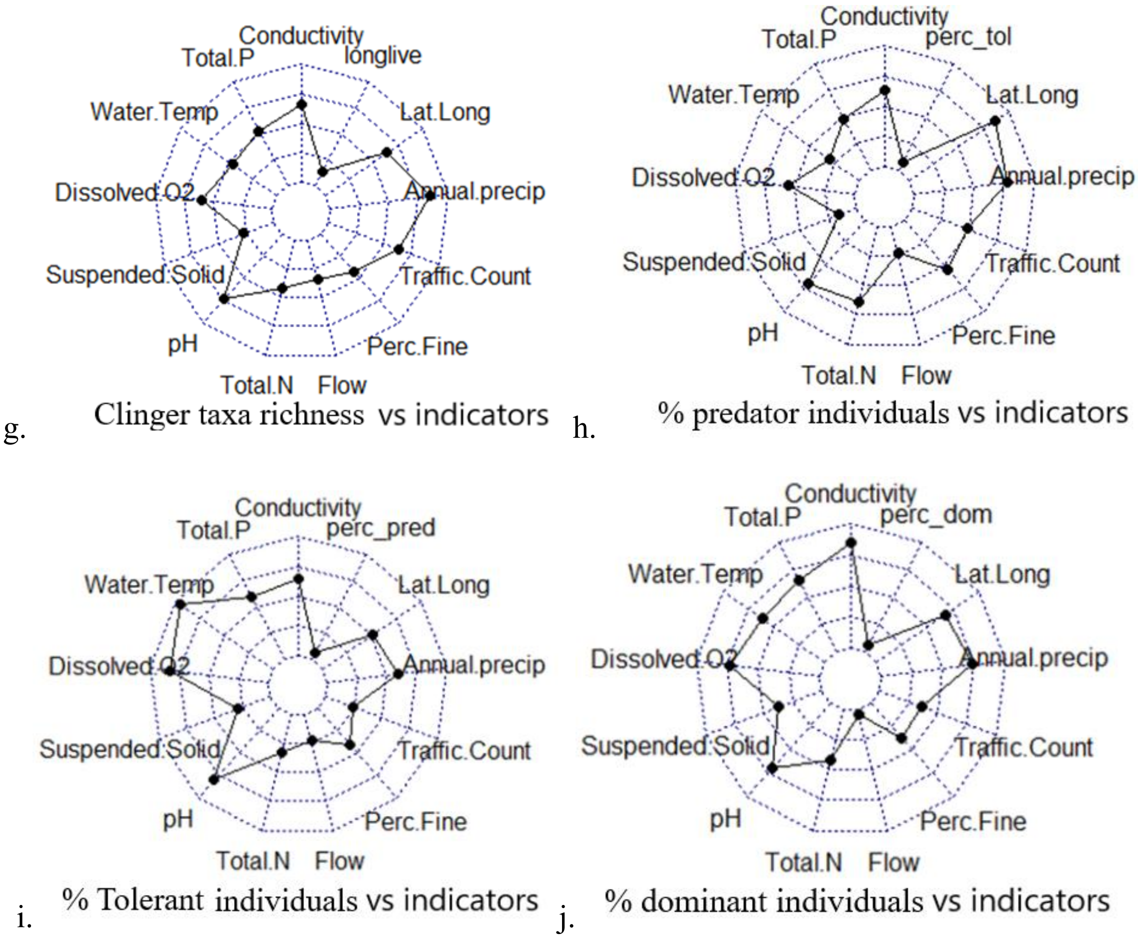
b. Ephemeroptera taxa vs indicators

c. Plecoptera taxa vs indicators

d. Trichoptera taxa vs indicators

e. Long-lived taxa vs indicators

f. Intolerant taxa vs indicators



g. Clinger taxa richness vs indicators

h. % predator individuals vs indicators

i. % Tolerant individuals vs indicators

j. % dominant individuals vs indicators

Figure 14. Result of Partial Least Square regression (PLS) expressing the relative influence of stream health indicators on individual B-IBI metrics in percentage, with all relative influence adding up to 100%. Each plot represented the relative influence of stream health indicator on one B-IBI metric, with (a) Total taxa richness, (b) Ephemeroptera taxa richness, (c) Plecoptera taxa richness, (d) Trichoptera taxa richness, (e) long-lived taxa, (f) intolerant taxa, (g) clinger taxa richness, (h) percent predator individuals, (i) percent tolerant individual, and (j) percent dominance.

3.3.4 Compiling WHI with BRT

The relationship between the WHI sub-indices and final WHI is presented in Table 3, each sub-index has a percent influence on the final WHI score. The intolerant species richness index has the highest influence at 30.50% and percent tolerant taxa index has the least influence on the final WHI at 2.14%.

Table 3. Percent influence of WHI sub-indices on final WHI score.

Sub-index	total	May	stone	caddis	longlive	intol	clinger	%pred	%tol	%dom
% influence	9.30%	7.44%	6.99%	4.96%	9.33%	30.50%	15.89%	7.16%	2.14%	6.24%

Table 4. Percent influence of individual indicators on WHI sub-indices.

Sub-index	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
total	14.41%	6.23%	4.58%	7.92%	14.19%	7.97%	5.40%	9.32%	6.69%	7.34%	3.35%	12.59%
may	9.28%	4.71%	30.50%	5.64%	2.39%	5.31%	12.10%	6.91%	4.80%	2.70%	2.41%	12.20%
stone	17.87%	12.23%	15.65%	5.59%	2.89%	5.21%	13.39%	7.59%	4.34%	4.14%	1.31%	9.72%
caddis	11.93%	8.11%	7.23%	12.35%	7.14%	8.18%	10.10%	9.12%	6.01%	6.83%	2.13%	10.81%
longlive	10.25%	7.72%	10.99%	8.04%	5.89%	8.17%	7.74%	12.22%	6.32%	7.32%	1.64%	13.65%
intol	9.93%	13.28%	20.61%	5.59%	2.56%	6.36%	15.85%	6.72%	4.56%	6.12%	2.00%	4.79%
clinger	11.26%	6.21%	7.15%	9.82%	8.83%	8.05%	7.62%	14.62%	7.33%	6.55%	1.92%	10.58%
%pred	9.01%	10.97%	13.05%	5.60%	2.03%	7.16%	7.05%	23.60%	2.57%	5.63%	2.54%	10.73%
%tol	18.84%	9.56%	9.12%	5.07%	9.50%	10.68%	6.67%	3.48%	5.10%	7.82%	1.13%	12.97%
%dom	10.04%	8.48%	9.37%	11.27%	9.17%	4.90%	6.95%	10.43%	6.65%	7.11%	1.92%	13.65%

In Table 4, the percemt influence of individual health indicators on each sub-index is presented. The percent influence of individual indicators adds up to 100% in each sub-index category. Precipitation has the least influence on the sub-indices overall, having the lowest percent influence on 9 out of the 10 sub-indices.

Table 5. Calculated percent influence of individual indicators on final WHI.

Indicators	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
% influence	11.32%	9.46%	14.31%	7.39%	5.71%	6.93%	10.69%	10.29%	5.43%	6.12%	2.08%	9.61%

Table 5 represents the percent influence of health indicators on final WHI score, result of compiling table 3 and table 4. Precipitation has the least influence on the final WHI at 2.08%, which is around 3% difference from the next lowest influencer.

3.3.5 Compiling WHI with PLS

The relationship between the WHI sub-indices and final WHI is presented in Table 6, each sub-index has a percent influence on the final WHI score. The percent predator index has the highest influence at 13.48% and percent tolerant taxa index has the least influence on the final WHI at 4.72%.

Table 6. Percent influence of WHI sub-indices on final WHI score.

Sub-index	total	May	stone	caddis	longlive	intol	clinger	%pred	%tol	%dom
% influence	9.45%	12.75%	9.41%	10.41%	10.24%	9.80%	12.87%	13.48%	4.72%	6.88%

Table 6 presents the percent influence of individual health indicators on each sub-index. The percent influence of individual indicators adds up to 100% in each sub-index category. Flow has the least influence on the sub-indices overall, having the lowest percent influence on 9 out of the 10 sub-indices.

Table 7. Percent influence of individual indicators on WHI sub-indices.

Sub-index	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
total	12.50%	9.93%	6.78%	12.57%	8.00%	11.20%	7.15%	1.18%	6.79%	4.89%	10.43%	8.60%
may	9.53%	9.44%	13.74%	9.87%	5.91%	10.33%	4.46%	3.71%	6.45%	6.19%	12.15%	9.02%
stone	10.96%	8.41%	10.00%	8.95%	4.44%	11.99%	4.75%	1.62%	7.04%	10.47%	11.63%	9.72%
caddis	12.49%	8.87%	4.49%	11.64%	4.26%	11.43%	7.84%	0.75%	5.27%	9.39%	12.20%	11.37%
longlive	10.03%	7.95%	6.88%	9.22%	4.09%	11.26%	6.44%	5.20%	6.56%	9.60%	13.07%	9.70%
intol	11.11%	9.14%	13.44%	10.41%	4.57%	11.64%	4.24%	2.36%	6.74%	6.12%	11.13%	9.09%
clinger	13.18%	8.57%	8.81%	12.01%	4.19%	11.07%	9.57%	1.51%	4.89%	3.27%	12.36%	10.57%
%pred	10.14%	9.20%	14.80%	12.83%	4.49%	12.53%	5.09%	3.52%	6.32%	3.81%	9.33%	7.93%
%tol	9.93%	7.62%	4.70%	8.73%	2.45%	10.89%	10.11%	3.56%	8.53%	7.61%	12.26%	13.62%
%dom	13.00%	10.02%	9.41%	11.22%	5.69%	10.78%	6.48%	0.52%	5.69%	5.50%	11.39%	10.30%

Results of Table 6 and Table 7 combined to make up the percent influence of health indicators on final WHI score, as shown in Table 5. Flow has the least influence on the final WHI at 2.45%, much lower than other indicators included in the study.

Table 8. Calculated percent influence of individual indicators on final WHI.

Indicators	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
% influence	11.27%	8.96%	9.82%	10.92%	4.89%	11.36%	6.42%	2.45%	6.28%	6.47%	11.54%	9.73%

3.3.6. Conversion of sample data to health score

The final step of constructing the WHI structures is to convert sampled watershed health indicator data into a health score for the individual indicators. The means, standard deviations, and the coefficient of variation of watershed health indicators of sites with B-IBI higher than 75% (healthy group) are calculated and shown in Table 9.

Table 9. The mean, standard deviation (SD) and the coefficient of variation (CV) of watershed health indicators calculated from the healthy group (sites with B-IBI > 75%). This table will be used to scale indicator samples to an indicator health score.

	Conductivity	Total P	Water Temp	Dissolved O2	Suspended Solid	pH	Total N	Flow	% Fine	latxlong	Traffic Count	Annual precipitation
Mean	102.38	0.0197	11.92	9.70	3.29	7.53	0.10	2.79	40.73	5706.40	4908.44	0.38
SD	84.30	0.0173	2.52	0.69	5.31	0.44	0.10	3.32	22.35	83.28	7812.40	0.36
CV	0.82	0.88	0.21	0.07	1.61	0.06	1.01	1.19	0.55	0.01	1.59	0.95

To convert the sampled indicator data collected from data sample sites, the means of each indicator are used as the 100% score thresholds, variation of more than 2 standard deviation from the mean results in a 0% score, and any variation less than 2 standard deviation from the mean will be scored linearly between 0% to 100%. As physical measurements of the indicators included cannot be lower than 0, the range for scaling cannot be less than 0. For instance, when scaling for traffic count, the range for scaling would be 0 passes/yr to 20532 passes/yr. Since 0 passes/yr is within 1 standard deviation, any sites with less than 4908 passes/yr (mean) will receive at least a 50% or higher score and considered to be in good condition.

The coefficient of variation of conductivity, total phosphorus, suspended solid, total nitrogen, stream flow, traffic count and annual precipitation was close or over 1.00, indicating high variance in the data. Indicators such as water temperature, dissolved oxygen, pH, percent fine sediments, and coordinate index (Lat x Long) had low coefficient of variation.

3.4. Discussion

3.4.1 Outcome of Study

The WHIs constructed in this study is meant to provide a data-based approach for watershed health assessment in the State of Washington, potentially providing local agencies a method that requires less resources for evaluation of local watersheds. We were able to construct WHIs using two models, the PLS and BRT model. While the WHIs presented different results, both models were able to identify important indicators that influence watershed health (IMR, 2004). The result from PLS in Table 10 shows that conductivity, dissolved oxygen, pH and annual precipitation to be the large influencers in watershed health. The WHI structure created with BRT, as shown in Table 10, indicated that conductivity and water temperature to be the larger drivers in watershed health. Both methods pointed to conductivity as the common indicator that has a high influence on watershed health, as conductivity is the only indicator that received a relative influence of more than 10% in both methods. Conductivity in stream water is known to be a strong indicator of external disturbance such as pollution, which may be the reason why both models captured conductivity as one of the large watershed health influencers (U.S. EPA, 2013). Dissolved oxygen, pH, and water temperature are all indicators that are also known to be influencers of freshwater system health (Walsh et al, 2001). In the Pacific Northwest, dissolved oxygen and water temperature heavily influence the recovery of the Pacific Salmon population (Oncorhynchus genus), which is a crucial part of the aquatic ecosystem (citation). The PLS and BRT models were able to recognize the critical indicators from the variance of the dataset.

The conversion for converting indicator measurement to indicator health score in section 3.3.6 (Table 9.) also presented interesting results. Indicators with low coefficient of variation in the conversion system seemed to be indicators that are sensitive to drastic changes, such as water temperature, dissolved oxygen, pH, and percent fine sediments. For instance, the scale of the conversion system for water pH was between pH 6.65 and pH 8.41 with pH 7.53 at 100% score, indicating a small range of variation. In healthy stream systems, the pH of water is unlikely to deviate much from pH 7 (Wetzel, 2001), which is a characteristic captured by the conversion system. The same conclusion could be drawn for other indicators as the conversion system was able to capture the indicators' sensitivity to changes.

Table 10. Comparison of percent influence of watershed health indicators on final WHI score between PLS and BRT.

Method	conductivity	Total P	Water Temp	Dissolved O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
PLS	11.27%	8.96%	9.82%	10.92%	4.89%	11.36%	6.42%	2.45%	6.28%	6.47%	11.54%	9.73%
BRT	11.32%	9.46%	14.31%	7.39%	5.71%	6.93%	10.69%	10.29%	5.43%	6.12%	2.08%	9.61%

While the PLS and BRT models captured the big drivers of watershed health, the WHI compiled with the two models (Table 10) resulted in different weight factors of individual indicators on the WHIs. The two models process the dataset in different ways (U.S. EPA, 2013; Khatri et al., 2020), which likely produced different results. Stream flow performed poorly as an indicator of watershed health in the PLS model, while having a relatively high influence in the BRT model. On the other hand, precipitation performed poorly in the BRT model but performed considerably better in the PLS model (2.08% in BRT and 11.54% in PLS). Our results showed that water temperature (BRT) and several water chemistry indicators (PLS) seemed to be larger influencers on the biological health of watershed. In BRT, precipitation data extracted from an external source seemed to perform poorly compared to other indicators.

With the WHI structures constructed, we randomly selected several sites and input the sampled data into our WHI structures to observe the outcome of our WHI compared to B-IBI. Our WHIs assess watershed health based on water quality related indicators, whereas the B-IBI score is a biological health assessment (Larson et al., 2019). From Table 11, we observed that the WHI scores and B-IBI scores show similar trend, sites with high B-IBI score received high WHI scores and vice versa. However, the B-IBI and WHI scores are not directly correlated to each other. We also noticed that the WHIs do not present extreme scores such as 0% or 100. This indicates that our WHIs and B-IBI are assessing different aspects of watershed health. While the B-IBI was used as the response variable for watershed health in our WHI creation process, the comparison between B-IBI and WHI results show that our WHIs are capable of capturing watershed health information that is not captured in B-IBI.

Table 11. Result of WHI score by entering site data back into WHI structures with PLS and BRT. Sites were randomly selected from original dataset.

Random Site	1	2	3	4	5	6
IBI Score	99.20%	98.79%	96.34%	21.56%	0.00%	5.03%
WHI score (PLS)	77.92%	85.12%	82.19%	19.02%	10.57%	23.41%
WHI score (BRT)	80.47%	82.34%	88.76%	10.81%	11.49%	9.07%

To further verify the functionality of the WHI structures, we repeated the process of creating the WHI structures with two sets of sample data extracted from the original dataset. The sample data with B-IBI scores higher than 75% and lower than 25% was separated into two sets of dataset for analysis. The radar graph in Figure 15 presents the WHI structures constructed in PLS with the separated datasets and the full dataset. The separated datasets resulted in a more similar WHI structure compared to the WHI structure from the original dataset.

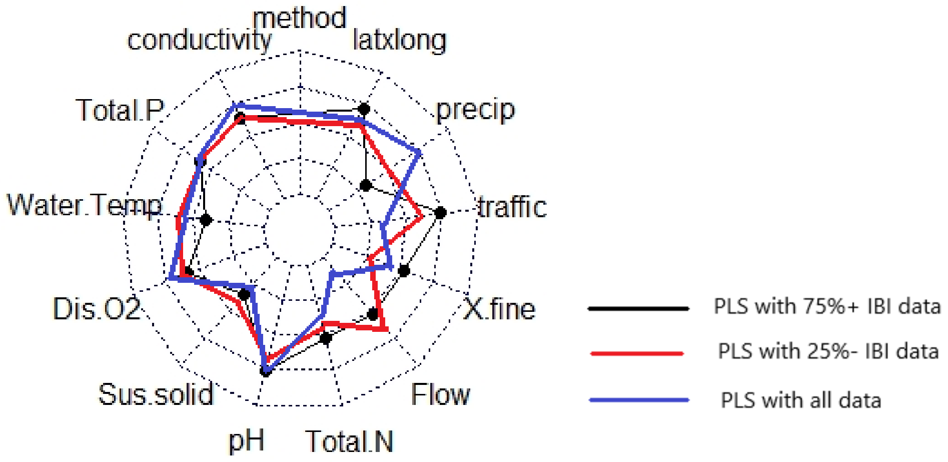


Figure 15. Radar graph of WHI structures (in relative influence) created with original dataset (blue), dataset with B-IBI scores higher than 75% (black), and dataset with B-IBI scores lower than 25% (red).

Constructing a WHI is a process of simplifying an extremely complex system in hopes of accurately predicting watershed health while minimizing resource use, which raises the potential of uncertainty. Our findings on large drivers of watershed health resulting from both the PLS and BRT models were not consistent with the findings by WADoE's previous relative risk assessment, which showed that percent fine sediment to be the biggest driver of stream biological health (Larson et al., 2019). The R^2 of most PLS regressions ranged from 0.05 to 0.30, which indicated that the models only captured a small portion of the variation in the dataset. There are several factors that potentially raised the uncertainty in the regressions, and some improvements can be made to reduce uncertainty in future studies.

The first potential factor is that the dataset from WA DoE was not complete, with some sample sites missing measurements for multiple indicators. While multivariate analyses can be powerful at handling large datasets, many multivariate analyses require complete numerical datasets (Khatri et al., 2020). There are methods to fill in missing data with imputation or interpolation (Wan et al.. 2020). However, the process of manipulating the dataset, while common, can result in higher uncertainties in methods such as PLS (Hao, 2020). For instance, the PLS regression (with B-IBI as the response variable, and B-IBI metrics as predictor variables) results presented in Table 12 shows that the dataset with a B-IBI score more than 75% and the dataset with B-IBI less than 25% show a different structure from the structure of original dataset. This should not be the case as the B-IBI is calculated from its metrics with predetermined equations and should have the same structure. We realized that the two datasets separated from the original dataset (the dataset with B-IBI higher than 75% and the dataset with B-IBI lower than 25%) did not include sample sites without B-IBI measurements, as the two separated datasets were separated by B-IBI measurement (B-IBI higher than 75% and lower than 25%) and sample sites without B-IBI measurements were not included. On the other hand, the original dataset contained imputed averages that replaced missing measurements for the possibility to be entered in a PLS regression model, which included all sample sites. Results in Table 12 shows that the two separated datasets had identical structures while the original dataset containing imputed averages showed a different structure. The PLS model differentiated the datasets without imputed data from the dataset with imputed data.

Table 12. The relative influence of B-IBI metrics on B-IBI scores resulted from PLS regression with original dataset, dataset with IBI > 75%, and dataset with IBI < 25%.

Dataset	total	May	stone	caddis	longlive	intol	clinger	%pred	%tol	%dom
Original dataset	9.45%	12.75%	9.41%	10.41%	10.24%	9.80%	12.87%	13.48%	4.72%	6.88%
Dataset with IBI >75%	7.23%	14.79%	5.68%	16.28%	7.04%	9.07%	15.52%	13.03%	3.64%	7.68%
Dataset with IBI < 25%	7.23%	14.79%	5.68%	16.28%	7.04%	9.07%	15.52%	13.03%	3.64%	7.68%

A challenge we faced during data analysis was the utilization of categorical data. The original dataset from the WADoE includes categorical data but is not used in the analysis of this study. The major limitation is that many multivariate analyses, such as the PLS model, can only take in continuous numerical data. Without a systematic method of converting categorical data into continuous numerical data, it is not possible to input such data into models like PLS (Hao, 2020). Another limitation of including categorical data is potentially introducing bias into the data during sample collection, such as judgement calls made by sampling personnel. The purpose of this study is to use a data-driven method that can eliminate as much human bias as possible, meaning that biased categorical data is not ideal for our purposes. To minimize human bias, it is best to avoid inclusion of categorical data and design data collection methods that maximize collection of continuous numerical data.

Another limitation with the dataset used in this study is that the data is collected in single collection events, rather than collected through data stations that monitor indicators continuously. Single collection events create a “snapshot” of the indicators, but it may not fully represent the current condition of the watershed. The “snapshot” data used in this study also lacks in quantity compared to datasets generated by data stations. This may explain the low influence of flow on final WHI in our results, as the sample may have captured extreme events that may not have been a good representation for the sites. While continuous data collection may require more resources to accomplish, it is the more ideal source of reliable large datasets. In future data collections for WHI construction, it would be best to include as much continuous numerical data as possible to increase accuracy of WHI scores.

We were able to fulfill our goal of creating a WHI for the State of Washington by using a data-driven approach, allowing the variation in datasets to describe the characteristics of water systems in the state. However, much is left to be explored as our analysis only explained a small portion of the variation in the data (R^2 of 0.05 to 0.30 in the PLS models). To maximize the extraction of information from the datasets, future studies will need to tailor data collection methods to the analytic tools accordingly. It is also crucial to explore new analytic tools for conducting watershed health related research, as new methods such as machine learning can potentially extract more information out of large datasets and help researchers understand the complex relationships between different systems in a watershed.

Chapter 4. Conclusion

Construction of WHI requires a large dataset of watershed health indicators that have minimum bias. Previous WHIs included indicators from at least 3 ecological attributes which encompassed a wide variety of watershed health indicators to aid in better watershed health assessment. However, there were assumptions made that may have unintentionally introduced bias. One assumption was that all sub-indices have equal influences on the final WHI score, which may not be true. Another assumption was that judgement-based data collection does not introduce bias and alter the study outcome. To minimize bias, the two assumptions made by previous WHIs must be addressed.

This study aimed to use a data-driven approach to create a WHI, with a goal of constructing an assessment system that can provide easier evaluation of watershed health in Washington State while minimizing potential bias. To minimize human bias, the data set was modified to only include direct measurements that did not involve any human-based judgments, such as the chemical concentration in sampled streams, before inputting the dataset into the statistical tools. Our analysis incorporated two multivariate analyses, the Partial Least Square Regression and Boosted Regression Tree, which provided us an insight into the ability of both methods when handling watershed health indicator data. The dataset included twelve variables for 1000+ sample events, resulting in high dimensionality of the dataset that required the use of analytic tools capable of dimension reduction. Multivariate analyses are known for their ability to overcome the problem of dimensionality in large multi-variable datasets (Khatri et al., 2020), which is the reason we chose to use multivariate models to explore our dataset. During data exploration, we used PCA biplot to get a glance of the dataset before moving onto supervised learning. The PCA biplot (Figure 10a.) presented relationships supported by current knowledge on benthic communities, such as positive relationship between B-IBI metric and dissolved oxygen level (Wetzel, 2001). Supervised learning models, such as PLS regression and BRT, provided us the ability to find relationships within large datasets that represent complex systems. We were able to construct a WHI based on the relationship between watershed health indicators and the health of the watersheds, which allowed us to simplify extremely complex watershed systems in order to evaluate watershed health. While some multivariate analytic methods are powerful, it is crucial to understand the data input requirements for each method. For instance, PLS regression and PCA have very low tolerance for incomplete datasets (Wan et al.. 2020), which can be difficult to use when there is missing data. Newer multivariate models, such as BRT, can tolerate a wide range of data type, including numerical and categorical (U.S. EPA, 2012). BRT is also insensitive to potential outliers or missing data in the dataset, and it does not assume linear relationships, making it capable of accounting for non-linear relationships between predictor and response variables (U.S. EPA, 2012). The ability to analyze categorical data can be a great advantage when working with environmental datasets, as complex systems such as watersheds cannot be measured only with numerical values.

We realized that the construction of WHI required a large dataset with continuous data and the inclusion of many watershed health indicator variables. During data exploration, we noticed that the data sampling events captured only a single moment of the conditions, or a “snapshot” of the indicators, which may not fully represent the actual condition of the sampled site. The “snapshot” data also lacks in quantity compared to datasets generated by data stations, where continuous, high quantity datasets are generated. While budget and resource limitations may constrain an organization from collecting more data, collaboration between different databases may potentially make up for the lack of data quantity and quality.

We were able to import external data from ArcGIS to build on top of our dataset for analysis, which may be a solution for increasing data quantity and quality. Maximizing the number of input variables during data exploration may also help with understanding the relationship between indicators and watershed health. Multivariate analytic tools are meant for handling high dimensionality datasets, which gives us the opportunity to explore the potential of using a wide variety of environmental variables. In this study, the watershed health indicator selection was limited by the availability of environmental variables in our dataset. However, future studies can potentially explore the use of other environmental variables, not used in this study, as watershed health indicators for constructing WHIs.

Utilization of complex analytic tools was crucial to the process of WHI construction, as we needed to process large datasets with high dimensionality. While multivariate analysis models used in this study were complex, we were able to use these models without much difficulty as easily accessible open-source software allowed us to integrate complex statistical models smoothly. Utilizing open-source tools can drastically reduce the time and resource needed to explore and manipulate datasets, which can aid agencies and scholars to process large datasets in a more efficient manner. Technological advancements are exponentially raising our ability to process datasets, which can potentially allow researchers to “digest” more datasets and extract relevant information. During data exploration, Self Organizing Maps (SOM) was explored as an option to create WHI. SOM is a new unsupervised learning using artificial neural network and has great potential for dimension reduction in large datasets. Machine learning and multivariate models can have great potential when constructing a WHI, with the assumption that the models are used appropriately. Prior statistical knowledge is an important factor when utilizing powerful statistical models, since without proper understanding, the outcome of the analysis will have minimal meaning. On the other hand, researchers also need to have prior knowledge or background information on the dataset being processed to appropriately extract information from the data. When handling environmental data, background knowledge on data is important as it could help data scientists spot potential trends or correlations during data processing. No prior background knowledge on the dataset can potentially result in lost of information, even when the appropriate analysis methods are used. Handling watershed data requires background knowledge on freshwater system as prior knowledge can help with maximizing efficiency and accuracy of data processing.

References

Ahn SR, Kim SJ. 2017. Assessment of integrated watershed health based on the natural environment, hydrology, water quality, and aquatic ecology. Hydrology and Earth System Sciences. 21(11):5583–5602.

Allan JD. 2004. Landscapes and Riverscapes: The Influence of Land Use on Stream Ecosystems. Annu Rev Ecol Evol Syst. 35(1):257–284. doi:10.1146/annurev.ecolsys.35.120202.110122.

City of Portland. 2019. Portland watershed health index summary. Bureau of Environmental Services.

Dudgeon D, Arthington AH, Gessner MO, Kawabata Z-I, Knowler DJ, Lévêque C, Naiman RJ, Prieur-Richard A-H, Soto D, Stiassny MLJ, et al. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol Rev Camb Philos Soc. 81(2):163–182. doi:10.1017/S1464793105006950.

Fleming B. 1999. Watershed health: An evaluation index for New Mexico. In: Finch, Deborah M; Whitney, Jeffrey C; Kelly, Jeffrey, F; Loftin, Samuel R Rio Grande ecosystems: linking land, water, and people: Toward a sustainable future for the Middle Rio Grande Basin; Albuquerque, NM Proc RMRS-P-7 Ogden, UT: US Department of Agriculture, Forest Service, Rocky Mountain Research Station 7:93–96.

Hao C. 2020. Comparison of partial least square algorithms in hierarchical laten variable model with missing data. Indexing & Metrics. 96(10):825-839. doi:0037549720944467.

Hascic I, Wu J. 2006. Land Use and Watershed Health in the United States. Land Economics. 82(2):214–239. doi:10.3368/le.82.2.214.

Henninger N, Revenga C, Brunner J, Payne R, Kassem K. 2000. Pilot analysis of global ecosystems: Freshwater systems. [accessed 2020 Dec 5]. https://www.wri.org/publication/pilot-analysis-global-ecosystems-2.

Hoque YM, Raj C, Hantush MM, Chaubey I, Govindaraju RS. 2014. How Do Land-Use and Climate Change Affect Watershed Health? A Scenario-Based Analysis. Water Qual Expo Health. 6(1):19–33. doi:10.1007/s12403-013-0102-6.

Hoque, Y.M., Tripathi, S., Hantush, M.M., Govindaraju, R.S., 2012. Watershed reli-ability, resilience and vulnerability analysis under uncertainty using waterquality data. J. Environ. Manag. 109, 101e11

Institute of Medicine (US) Roundtable on Environmental Health Sciences R, Goldman L, Coussens CM. 2004. Overview of Environmental Health Monitoring and the Use of Indicators. National Academies Press (US). [accessed 2021 Aug 11]. https://www.ncbi.nlm.nih.gov/books/NBK215456/.

Karim Bengraïne, Taha F Marhaba, Using principal component analysis to monitor spatial and temporal changes in water quality, Journal of Hazardous Materials, Volume 100, Issues 1–3, 2003, Pages 179-195, ISSN 0304-3894.
https://doi.org/10.1016/S0304-3894(03)00104-3.

Khatri, P., Gupta, K.K. & Gupta, R.K. 2020. A review of partial least squares modeling (PLSM) for water quality analysis. Model. Earth Syst. Environ.
https://doi.org/10.1007/s40808-020-00995-4

Larson CA, Merritt G, Janisch J, Lemmon J, Rosewood-Thurman M, Engeness B, Polkowske S, Onwumere G. 2019. The first statewide stream macroinvertebrate bioassessment in Washington State with a relative risk and attributable risk analysis for multiple stressors. Ecological Indicators. 102:175–185. doi:10.1016/j.ecolind.2019.02.032.

Maddock I. 2001. The Importance of Physical Habitat Assessment for Evaluating River Health. Freshwater Biology. 1999; 41 (2) : 373-391.

Mimikou MA, Baltas E, Varanou E, Pantazis K. 2000. Regional Impacts of Climate Change on Water Resources Quantity and Quality Indicators. Journal of Hydrology. 234 (1-2): 95-109.

Minnesota Department of Natural Resources. Watershed Health Assessment Framework, Health scores. https://www.dnr.state.mn.us/whaf/about/scores/index.html. Retrieved on January 4th, 2021

Momenian P, Nazarnejhad H, Miryaghoubzadeh M, Mostafazadeh R. Assessment and Prioritizing of Subwatersheds Based on Watershed Health Scores (Case Study: Ghotorchay, Khoy, West Azerbaijan). jwmr. 2018; 9 (17) :1-13.

Peters NE, Meybeck M. 2000. Water Quality Degradation Effects on Freshwater Availability: Impacts of Human Activities. Water International. 25(2):185–193. doi:10.1080/02508060008686817.

Rice J. 2003. Environmental health indicators, Ocean & Coastal Management. 46(3):235-259. Doi:10.1016/S0964-5691(03)00006-1.

Robert I. McDonalda,1, Pamela Greenb, Deborah Balkc, Balazs M. Feketeb, Carmen Revengaa, Megan Toddc, and Mark Montgomeryd. 2011. Urban growth, climate change, and freshwater availability.

Rosenberg DM, Resh VH. 1993. Freshwater biomonitoring and benthic macroinvertebrates. New York: Chapman & Hall.

Rodell, M., Famiglietti, J.S., Wiese, D.N. et al. Emerging trends in global freshwater availability. Nature 557, 651–659 (2018). https://doi.org/10.1038/s41586-018-0123-1

Shuster WD, Bonta J, Thurston H, Warnemuende E, Smith DR. 2005. Impacts of impervious surface on watershed hydrology: A review. Urban Water Journal. 2(4):263–275. doi:10.1080/15730620500386529.

Tanaka MO, Souza ALT, Moschini LE, Oliveria AK. 2016. Influence of Watershed Land Use and Riparian Characteristics on Biological Indicators of Stream Water Quality in Southeastern Brazil. Agriculture, Ecosystems & Environment. 216:333-339.

Tirkey P, Bhattacharya T, Chakraborty S. 2013. Water quality indices- important tools for water quality assessment: a review. 1.

US EPA. 2012. Identifying and Protecting Healthy Watersheds: Concepts, Assessments, and Management Approaches. U.S. Environmental Protection Agency.

U.S. EPA. 2013. California Integrated Assessment of Watershed Health. https://www.mywaterquality.ca.gov/monitoring_council/healthy_streams/docs/ca_hw_report_111213.pdf.

Walsh CJ, Roy AH, Feminella JW, Cottingham PD, Groffman PM, Morgan RP. 2005. The urban stream syndrome: current knowledge and the search for a cure. jnbs. 24(3):706–723. doi:10.1899/04-028.1.

Wetzel RG. 1992. Clean water: a fading resource. Hydrobiologia. 243(1):21–30. doi:10.1007/BF00007017.

Wetzel RG. 2001. Limnology: lake and river ecosystems. [accessed 2020 Dec 3]. http://site.ebrary.com/id/10606261.

WA DNRP. 2014. Identifying Stressor Risk to Biological Health in Streams and Small Rivers of Western Washington. Washington Department of Natural Resources and Parks

Wan SY, Agus S, Perumal K. 2020. Missing data treatment for locally weighted partial least square-based modelling: A comparative study. Chemical Engineering. 15(2).

Appendix A: PLS regression for samples with B-IBI > 75

IBI Metrics	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
total	8.12%	9.44%	8.70%	8.40%	7.96%	10.99%	7.21%	6.78%	8.07%	10.60%	3.31%	10.46%
may	9.77%	8.60%	10.64%	11.10%	2.01%	10.25%	8.83%	6.78%	7.79%	10.80%	3.34%	10.00%
stone	9.74%	8.92%	1.46%	10.23%	2.55%	13.04%	10.21%	2.64%	7.31%	12.82%	8.70%	12.32%
caddis	10.64%	7.26%	2.76%	7.44%	8.26%	12.10%	4.20%	12.08%	7.95%	12.20%	5.42%	9.62%
longlive	10.55%	9.11%	8.71%	5.86%	7.85%	10.14%	6.89%	3.63%	9.63%	13.12%	3.00%	11.44%
intol	8.33%	10.20%	7.60%	6.41%	3.36%	11.15%	4.77%	11.54%	6.55%	10.56%	9.26%	10.23%
clinger	7.64%	8.02%	3.36%	9.95%	6.20%	11.20%	9.76%	9.37%	8.01%	11.11%	3.39%	11.82%
%pred	11.15%	10.55%	6.46%	8.43%	4.01%	11.45%	8.56%	7.63%	8.19%	9.93%	2.82%	10.75%
%tol	10.84%	8.02%	4.65%	10.10%	1.52%	11.22%	13.37%	0.96%	7.67%	11.72%	8.44%	11.44%
%dom	7.90%	10.66%	8.11%	11.00%	5.13%	12.35%	7.71%	6.74%	8.57%	9.25%	2.34%	10.17%

Appendix B: PLS regression for samples with B-IBI < 25

IBI Metrics	conductivity	Total P	Water Temp	Dis O2	Sus solid	pH	Total N	Flow	%fine	traffic	precip	latxlong
total	10.69%	9.44%	10.20%	8.75%	9.07%	11.03%	8.57%	10.41%	1.64%	5.94%	5.60%	8.54%
may	10.97%	11.16%	5.18%	5.60%	7.47%	10.46%	7.76%	10.43%	4.10%	11.67%	5.50%	9.64%
stone	9.04%	14.00%	3.22%	4.40%	11.81%	10.17%	10.92%	10.16%	5.02%	7.71%	4.21%	9.27%
caddis	5.99%	8.08%	13.03%	8.81%	10.07%	8.83%	3.85%	11.92%	3.53%	6.29%	8.48%	11.06%
longlive	9.45%	11.58%	8.50%	11.52%	7.46%	9.38%	3.15%	10.49%	3.79%	13.18%	4.78%	6.66%
intol	10.83%	9.84%	9.22%	11.47%	10.75%	11.45%	7.19%	6.68%	2.40%	5.61%	6.44%	8.06%
clinger	10.84%	3.39%	14.12%	11.47%	2.28%	13.00%	5.94%	11.84%	2.27%	11.24%	9.54%	4.00%
%pred	9.91%	8.60%	11.19%	11.63%	6.31%	9.89%	10.96%	10.64%	2.94%	3.87%	5.79%	8.21%
%tol	9.73%	11.21%	10.12%	9.36%	12.16%	10.77%	7.37%	2.23%	4.19%	4.09%	9.58%	9.12%
%dom	11.61%	9.50%	9.12%	9.17%	6.03%	9.98%	3.60%	2.64%	7.03%	11.82%	9.81%	9.62%

Appendix C: Python Code for Data Extraction

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import glob

biodata = pd.read_csv(‘#source folder#¥¥Bio data in ORDER1.csv', index_col = None)
bioID = pd.read_csv('#source folder#¥¥bio genus id.csv', index_col = None)

j = []
totalscore = [0]
stonecount = [0]
maycount = [0]
caddiscount = [0]
i3 = [] #for naming final macro counts
macrocount = []

for i in enumerate(biodata["Site_ID"]) :
    i1 = i[0]
    i2 = i[1]
    if i2 == j :
        taxalev = biodata.iloc[i1]["Result_Taxon_Level"]
        if taxalev == "GENUS" :
            name = biodata.iloc[i1]["Scientific_Name"]
            for x1 in enumerate(bioID["Genus"]):
                idnum = x1[0]
                idname = x1[1]
                if idname == name :
                    idorder = bioID.iloc[idnum]["Order"]
                    if idorder == "Plecoptera" :
                        singlecount = biodata.iloc[i1]["Result_Value"]
                        stonecount = stonecount + singlecount
                    if idorder == "Ephemeroptera" :
                        singlecount = biodata.iloc[i1]["Result_Value"]
                        maycount = maycount + singlecount
                    if idorder == "Trichoptera" :
                        singlecount = biodata.iloc[i1]["Result_Value"]
                        caddiscount = caddiscount + singlecount
    else :
        sitecount = [i3, stonecount, maycount, caddiscount]
        macrocount.append(sitecount)
        stonecount = [0]
        maycount = [0]
        caddiscount = [0]
    j = i2
    i3 = i[1]

pd.DataFrame(macrocount)

chemdata = pd.read_csv('#source folder#¥¥Chem data.csv', index_col = None)

j = []
chemcompile = []
i3 = [] #for naming final macro counts
conduc = 0
totalP = 0
watertemp = 0
disso2 = 0
suspsolid = 0
phvalue = 0
totalN = 0


for i in enumerate(chemdata["Data_Collection_Event"]) :
    i1 = i[0]
    i2 = i[1]
    if i2 == j :
        chempara = chemdata.iloc[i1]["Parameter_Name"]
        chemresult = chemdata.iloc[i1]["Result_Value"]
```

```
if chempara == "Conductivity" :
    if chemresult > 300 :
        conduc = 1
    if chemresult <= 300 and chemresult >= 200:
        conduc = 2
    if chemresult < 200 :
        conduc = 3
if chempara == "Total Phosphorus" :
    if chemresult > 0.036 :
        totalP = 1
    if chemresult <= 0.036 and chemresult >= 0.014:
        totalP = 2
    if chemresult < 0.014 :
        totalP = 3
if chempara == "Temperature, Water" :
    if chemresult > 20 :
        watertemp = 1
    if chemresult <= 20 and chemresult >= 12:
        watertemp = 2
    if chemresult < 12 :
        watertemp = 3
if chempara == "Dissolved Oxygen" :
    if chemresult > 7 :
        disso2 = 1
    if chemresult <= 7 and chemresult >= 10 :
        disso2 = 2
    if chemresult < 10 :
        disso2 = 3
if chempara == "Total Suspended Solids" :
    if chemresult > 44 :
        suspsolid = 1
    if chemresult <= 44 and chemresult >= 2 :
        suspsolid = 2
    if chemresult < 2 :
        suspsolid = 3
        if chempara == "pH" :
            if chemresult < 6.5 or chemresult > 8.5 :
                phvalue = 1
            if chemresult < 7.0 or chemresult > 7.5    :
                phvalue = 2
            if chemresult >=7.0 and chemresult <= 7.5 :
                phvalue = 3
        if chempara == "Total Persulfate Nitrogen" :
            if chemresult > 0.462    :
                totalN = 1
            if chemresult <= 0.462 and chemresult >= 0.131    :
                totalN = 2
            if chemresult < 0.131 :
                totalN = 3

    else :
        sitedata = [i3, conduc, totalP, watertemp, disso2, suspsolid, phvalue, totalN]
        chemcompile.append(sitedata)
        conduc = 0
        totalP = 0
        watertemp = 0
        disso2 = 0
        suspsolid = 0
        phvalue = 0
        totalN = 0

    j = i2
    i3 = i[1]

chemcompile_df = pd.DataFrame(chemcompile, columns = ['id', 'Conductivity', 'Total P',
'Water Temp', 'Dissolved O2','Suspended Solid','pH','Total N'])
```

Appendix D: Python code for data normalization and missing data imputation

```
dataset = pd.read_csv('#source folder#¥¥extracted_dataset.csv', index_col = 0)
y = dataset.values
x = SimpleImputer(missing_values=np.nan, strategy='mean')
x.fit(y)
newdata = x.transform(y)
x = newdata
x = StandardScaler().fit_transform(x)
np.mean(x),np.std(x)
feat_cols=['total','may','stone','caddis','longlive','intolerant','clinger','perc_pred','perc_tol','pe
rc_dom','Conductivity','TotalP','Water Temp','Dissolved O2','Suspended Solid','pH','Total
N','Flow','% Fine','Traffic','Annual Precip','Lat.Long']
normalized_data = pd.DataFrame(x,columns=feat_cols)
```

Appendix E: R code for PLS analysis

```
setwd("##root##/New data")
getwd()
data <-read.csv("Merged no outlier impute data.csv")
library(plsdepot)
set.seed(1000)

## With IBI vs IBI metric
xvar = data[c(2)]
pls_model2= plsreg1(data[,3:12], xvar, comps=4, crosval = TRUE)
plot(pls_model2)
pls_model2$cor.xt
pls_model2$x.loads
pls_model2$expvar
pls_model2$Q2
pls_model2$Q2cum
pls_model2$reg.coefs
pls_model2$std.coefs
pls_model2$R2Xy
sum1 = sum(pls_model2$R2Xy[1:10,3])
IBIR2 = (pls_model2$R2Xy[1:10,3])/sum1
pls_R2 = as.data.frame(pls_model2$expvar)
plot(pls_R2$R2Ycum,xlab="# of components", ylab="R^2 Y cumulative")
data_1 =    as.data.frame(t(pls_model2$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,11),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,11),data_3)
```

```
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,11),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,11),data_5)
radarchart(plotdata3)


## With May
pls_model = plsreg1(data[,13:24], data[,4,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
MayR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)
```

```
## Total
pls_model = plsreg1(data[,13:24], data[,3,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
TotalR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)




## Stone
pls_model = plsreg1(data[,13:24], data[,5,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
```

```
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
StoneR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Caddis
pls_model = plsreg1(data[,13:24], data[,6,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
CadR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
```

```
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Longlive
pls_model = plsreg1(data[,13:24], data[,7,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
longliveR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
```

```
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Intolerant
pls_model = plsreg1(data[,13:24], data[,8,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
intolR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
```

```
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Clinger
pls_model = plsreg1(data[,13:24], data[,9,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
clingerR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Percent Predator
pls_model = plsreg1(data[,13:24], data[,10,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
```

```
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
percpredR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Percent Tolerant
pls_model = plsreg1(data[,13:24], data[,11,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
perctolR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
```

```
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)

## Percent Dominance
pls_model = plsreg1(data[,13:24], data[,12,drop=FALSE], comps=5, crosval = TRUE)
plot(pls_model)
pls_model$cor.xyt
pls_model$reg.coefs
pls_model$R2Xy
sum1 = sum(pls_model$R2Xy[1:12,5])
percdomR2 = (pls_model$R2Xy[1:12,5])/sum1
data_1 =    as.data.frame(t(pls_model$R2Xy))
data_2 = data_1[1,]
data_3 = data_1[2,]
data_4 = data_1[3,]
data_5 = data_1[4,]
data_6 = data_1[5,]

library(fmsb)
plotmax = c(1,1,1,1,1,1,1,1,1,1,1,1,1)
plotdata <- rbind(plotmax,rep(0,13),data_2)
```

```
radarchart(plotdata)
plotdata1 <- rbind(plotmax,rep(0,13),data_3)
radarchart(plotdata1)
plotdata2 <- rbind(plotmax,rep(0,13),data_4)
radarchart(plotdata2)
plotdata3 <- rbind(plotmax,rep(0,13),data_5)
radarchart(plotdata3)
plotdata4 <- rbind(plotmax,rep(0,13),data_6)
radarchart(plotdata4)
```

Appendix F: R code for BRT analysis

```
setwd("##root##/New data")
getwd()
data1 <-read.csv("processed data.csv")
library(dismo)
library(tidyverse)
## IBI vs metrics
BRT_model <- gbm.step(data = data1, gbm.x = 3:12, gbm.y= 2, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)
BRT_1 = as.data.frame(t(summary(BRT_model)))
BRT_result = BRT_1[2,]

## total vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=3, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

## May vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=4, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##Stone vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=5, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##Caddis vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=6, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##longlive vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=7, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##intolerant vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=8, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##clinger vs indicators
```

```
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=9, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##perc_pred vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=10, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##perc_tol vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=11, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)

##perc_dom vs indicators
BRT_model <- gbm.step(data = data1, gbm.x = 13:24, gbm.y=12, family = "gaussian",
tree.complexity = 5, learning.rate=0.01, bag.fraction = 0.5)
names(BRT_model)
summary(BRT_model)
```